

Rosenbrock-Type Methods for Semilinear Parabolic Equations

Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von
Leon Schramm

Kiel, 2019

Erster Gutachter: Prof. Dr. Malte Braack

Zweiter Gutachter: Prof. Dr. Jens Lang

Tag der mündlichen Prüfung: 4.11.2019

Abstract

In this work we examine the viability of Rosenbrock-type time-stepping methods — specifically Rosenbrock-Wanner (ROW) methods and W-methods — for the temporal discretization of certain parabolic partial differential equations (PDEs) that have a dominating linear term (usually the Laplacian) and lower order nonlinearities, such as convective or reactive terms.

The original aim when starting the research on this subject was to improve upon standard ways to discretize in time the incompressible Navier-Stokes equations. For stability reasons, Runge-Kutta methods applied to those equations have to be *implicit* — leading to the numerically expensive task of solving large nonlinear systems of equations. Rosenbrock-type methods are closely related to implicit Runge-Kutta methods but avoid the handling of nonlinear systems by treating only the *linearized* approximation of a nonlinear equation implicitly. Whether this design maintains stability of the time-stepping scheme depends on the type of the nonlinear equation and the way it is linearized. The advantage is that — compared to implicit Runge-Kutta methods — the stiffness matrix has to be build less often and fewer linear systems have to be solved.

In a two-way approach we investigate the application of Rosenbrock-type methods to parabolic PDEs both in theory and in numerical experiments.

For the theoretical studies we restrict ourselves to the *semi*-discretization in time by Rosenbrock-type methods, which is usually described in a framework of semilinear parabolic equations, sectorial operators and semigroups. Since *partial* differential equations which fit into that framework may be handled very similarly to *ordinary* differential equations (ODEs), we first introduce Rosenbrock-type methods through their application to ODEs. We then demonstrate that certain types of convection-diffusion-reaction equations, certain Oseen-type equations, and the incompressible Navier-Stokes equations do fit into the above mentioned framework. We also show that some existing semi-discrete (in time) convergence results for Rosenbrock-type methods may be applied to these equations — with the W-method results requiring extra effort because those methods allow the user considerable freedom in choosing how to linearize a nonlinear parabolic PDE.

In this work the usefulness of Rosenbrock-type methods and the above mentioned semi-discrete error estimates is checked by conducting numerical experiments with fine spatial discretizations and chosen exact solutions that are at each point in time approximated *exactly* by the spatial discretization. Using this approach, we numerically test the application of various ROW methods and W-methods to a convection-diffusion problem, a reaction-diffusion problem and an incompressible Navier-Stokes problem.

Lastly, we also numerically examine a more practical problem setup without known exact solution. We opt for a well-studied variant of the famous benchmark flow around a circular obstacle. The performance of our Rosenbrock-type methods is assessed by comparing lift, drag and pressure difference of the numerical solutions with a reliable reference solution from the literature.

Zusammenfassung

In dieser Arbeit untersuchen wir die Eignung von Rosenbrock-Typ-Methoden (RTM) — speziell von Rosenbrock-Wanner- (ROW) und W-Methoden — für die zeitliche Diskretisierung von bestimmten parabolischen partiellen Differentialgleichungen (PDG), die einen dominierenden linearen Term haben (z. B. den Laplace-Operator) und Nichtlinearitäten niedrigerer Ordnung, wie z. B. konvektive oder reaktive Terme.

Unser ursprüngliches Ziel war die Verbesserung von Standardmethoden zur Zeitdiskretisierung der inkompressiblen Navier-Stokes-Gleichungen. Aus Stabilitätsgründen sollten Runge-Kutta-Verfahren (RKV), welche auf diese Gleichungen angewandt werden, *implizit* sein — das führt aber zu der numerisch teuren Aufgabe, große, nichtlineare Gleichungssysteme (GLS) lösen zu müssen. RTM sind eng verwandt mit impliziten RKV, umgehen aber nichtlineare GLS dadurch, dass sie nur die *linearisierte* Approximation einer nichtlinearen Gleichung implizit behandeln. Ob solch eine Methode stabil ist, hängt von der nichtlinearen Gleichung und der speziellen Linearisierung ab. Der praktische Vorteil im Vergleich zu impliziten RKV ist, dass die Steifigkeitsmatrix weniger oft aufgebaut werden muss und dass weniger lineare GLS gelöst werden müssen.

In einem zweigleisigen Ansatz untersuchen wir die Anwendung von RTM auf parabolische PDG sowohl theoretisch als auch in numerischen Experimenten.

Für die Theorie beschränken wir uns auf die *Semi*-Diskretisierung in der Zeit durch RTM, welche gewöhnlich in einem Framework von semilinearen parabolischen Gleichungen, sektoriellen Operatoren und Halbgruppen beschrieben wird. Da *partielle* Differentialgleichungen, die in dieses Framework passen, sehr ähnlich wie *gewöhnliche* Differentialgleichungen (GDG) behandelt werden können, führen wir RTM zunächst anhand ihrer Anwendung auf GDG ein. Dann zeigen wir, dass bestimmte Konvektions-Diffusions-Reaktions- und Oseen-Typ-Gleichungen sowie die inkompressiblen Navier-Stokes-Gleichungen in das genannte Framework passen. Außerdem zeigen wir die Anwendbarkeit einiger existierender semi-diskreter Konvergenzresultate für RTM auf diese Gleichungen — dabei erfordern die W-Methoden am meisten Anstrengung, da sie beträchtlichen Spielraum bei der Wahl der Linearisierung einer nichtlinearen Gleichung lassen.

Wir testen die Nützlichkeit von RTM und der genannten semi-diskreten Konvergenzresultate durch numerische Experimente mit feinen örtlichen Diskretisierungen und gewählten exakten Lösungen, die *exakt* von der örtlichen Diskretisierung approximiert werden. Wir untersuchen so die Anwendung einiger ROW- und W-Methoden auf ein Konvektions-Diffusions-Problem, ein Reaktions-Diffusions-Problem und ein inkompressibles Navier-Stokes-Problem.

Zuletzt führen wir auch noch Experimente mit einem praktischeren Problem durch, bei welchem die Lösung nicht bekannt ist. Wir wählen eine gut erforschte Variante der bekannten Benchmark-Strömung um ein kreisförmiges Hindernis. Die Performance unserer Methoden wird dabei eingeschätzt mit Hilfe von Drag-, Lift- und Druckdifferenzwerten der numerischen Lösung, die wir mit verlässlichen Referenzwerten aus der Literatur vergleichen.

Contents

1	Introduction	1
2	Physical Background	7
2.1	Mass and Heat transfer	8
2.2	Incompressible Flow	9
3	Semilinear Parabolic Model	11
3.1	Basic Notations and Mathematical Concepts	11
3.2	Sectorial Operators	19
3.3	Convection-Diffusion Operator	23
3.4	Stokes Operator	29
3.5	Fractional Spaces	36
3.5.1	Examples of Fractional Spaces	41
3.6	Extended Operators	44
3.6.1	Examples of Extended Operators	47
3.7	Problem Statement for Semilinear Parabolic Equations (SPEs)	49
3.8	The Incompressible Navier-Stokes Equations as an SPE	53
3.8.1	The Oseen Operator	59
4	Discretization in Time by Rosenbrock-Type Methods	66
4.1	Rothe's Method	67
4.2	Rosenbrock-Type Methods for Ordinary Differential Equations	68
4.3	Order, Stability and Dissipation of ODE solvers	72
4.4	Order Reduction and More Accurate Concepts of Convergence for Stiff ODEs	79
4.5	ROW Methods and W-Methods for SPEs	88
4.5.1	Application of ROW Methods and W-Methods to Some Specific SPEs	94
4.5.2	Smoothness Assumptions and Order Reduction	102
4.6	Specific Rosenbrock-type Methods with Parameter Tables	104
5	Discretization in Space by the Finite Element Method	107
5.1	P_r - and Q_r -Elements	109
5.1.1	P_r - and Q_r -Elements for the Stationary Convection-Diffusion Equation	110
5.1.2	Q_r -Elements for an Oseen-Type Equation	112
5.2	Notes on the Full Discretization	115
6	Numerical Experiments	118
6.1	A Convection-Diffusion Problem with Known Solution	119
6.2	A Reaction-Diffusion Problem with Known Solution	131
6.3	A Navier-Stokes Problem with Known Solution	141
6.4	Benchmark Flow Around a Circular Obstacle	164

7 Summary, Conclusion and Outlook	176
List of Figures	180
List of Tables	182
Bibliography	186

1 Introduction

The motivation for starting this work was to provide a better understanding of Rosenbrock-type time-stepping methods — in particular the so-called W-methods — when they are applied to the incompressible Navier-Stokes equations. It is well-known that time-stepping schemes for the Navier-Stokes equations usually have to be implicit for stability reasons (see [12] and the introduction of [32]). On the other hand, in implicit schemes the nonlinearity of these equations leads to nonlinear systems, the solution of which is often numerically expensive. By their design, Rosenbrock-type methods (see [56] for Rosenbrock’s initial idea), such as the Rosenbrock-Wanner (ROW) and W-methods, promise to be well-suited in this situation as in some sense they “reduce implicitness” as much as possible while still maintaining sufficient stability *if* the equation has the appropriate structure.

Roughly speaking, if an evolution equation has a dominating linear part which “contains most of the stiffness” — such as the Laplacian — and lower order but possibly nonlinear parts — such as convective terms — then the Rosenbrock-type methods will keep stability without it being necessary to solve nonlinear equations. This is achieved in those methods by treating the stiff linearization of the equation implicitly and handling the nonlinear error terms explicitly. In practice, the numerical speedup then mainly stems from fewer assemblies of the stiffness matrix and fewer linear systems to be solved.

W-methods (first introduced by Steihaug and Wolfbrandt in [62]) are extreme in this regard, as in contrast to ROW methods (initial publications include [25], [72] and [34]), they allow the user to linearize a nonlinear equation by only *approximating* the exact Fréchet derivative. Instead of the exact Fréchet derivative for the current time step, one could use the Fréchet derivative from a previous time step — in some cases one might even be able to just use the same approximate Fréchet derivative at *every* time step, meaning that building only *one* stiffness matrix during the entire time-stepping process could be enough.

The idea behind Rosenbrock-type methods becomes less appealing when an equation does not have the above described structure — for very large Reynolds numbers, the Navier-Stokes equations for example are generally not dominated by the Laplacian anymore. In such a situation one would certainly worry about loss of accuracy and/or stability of the numerical schemes when treating only the Fréchet derivative of the nonlinearity implicitly (in ROW methods and W-methods) or even just approximating that Fréchet derivative (in W-methods).

Since the focus of this work lies on the temporal discretization of parabolic partial differential equations (PDEs), we concentrate on the time derivative by combining the spatial derivatives into an unbounded operator between Sobolev spaces. In an abstract framework of sectorial operators and semigroups, we can then handle our *partial* differential equations similarly to *ordinary* differential equations (ODEs). This lets us conveniently formulate and examine the semi-discretization in time of parabolic PDEs by employing notation, methods and techniques

from the field of ODEs. The resulting semi-discrete error estimates hold independently of any specific spatial discretization that might be used in practice and can be seen as covering the situation that in a fully-discrete algorithm the spatial mesh size of a suitable spatial discretization nears zero. Roughly speaking, the temporal stiffness of an equation is completely represented in the semi-discrete error estimates.

Another reason for this abstract approach is that a substantial part of the literature for Rosenbrock-type methods applied to parabolic partial differential equations is formulated in this framework. First and foremost, we want to make use of a series of papers by Lubich and Ostermann [40, 42, 41], in which the authors prove sharp temporal error estimates for Runge-Kutta methods, ROW methods and W-methods that are applied to certain types of abstract evolution equations in Hilbert spaces.

At this point, we broaden our initial scope of discretizing in time the incompressible Navier-Stokes and other parabolic equations, such as the convection-diffusion equation. A large part of this work concerns itself with showing that these equations are what we call semilinear parabolic equations (SPE) and that they satisfy all further requirements posed in the convergence results by Ostermann and Lubich. Of course, most of this is known — see for instance [18] and example 3.8 in [27] for examinations of the incompressible Navier-Stokes equations in this regard. On the other hand, classifications of Oseen-type operators as sectorial operators or the incompressible Navier-Stokes equations as an SPE are often not thoroughly explained.

We define a convection-diffusion operator as the Dirichlet Laplacian perturbed by lower order terms and, similarly, we define an Oseen operator as the Stokes operator perturbed by lower order terms. Using extra care to handle the loss of symmetry introduced by any convective terms, we then provide a comprehensive classification of our convection-diffusion operator, the Stokes operator and our Oseen operator as sectorial operators. Albeit not using new concepts, our examination of the Oseen operator is not based on any literature for that specific topic. We also show in detail (by carefully reiterating the arguments from example 3.8 in [27]) that the incompressible Navier-Stokes equations can indeed be formulated as an SPE — naturally, the investigation of the nonlinearity is essential in that demonstration.

When verifying that the application of Rosenbrock-type methods to the above mentioned equations is covered by the convergence theory, W-methods are of particular interest. These methods demand extra scrutiny because algorithmically they allow *arbitrary* operators in place of the *exact* Fréchet derivatives (evaluated at the numerical solution), which would be used in ROW methods. To be able to prove convergence of the semi-discretization in time, however, these approximate operators should not be completely arbitrary. We explore a few options for these operators and show in detail that they fulfill the requirements posed in the convergence theory.

To the author's knowledge, an examination of possible approximate operator choices in W-methods which are used to semi-discretize in time the two or three dimensional incompressible Navier-Stokes equations, had not been published before. Of note here is the paper [61] by Schwitzer, in which the author applies W-methods to Burger's equation by approximating the full Fréchet derivative by just its principal part — a straightforward and sensible choice which we also discuss for the Navier-Stokes equations.

As opposed to general W-methods, the application of ROW methods to various parabolic problems, such as the Navier-Stokes problem, has been extensively studied. In the lecture notes [37]

by Lang and the papers [38, 67], this has been done in a similar framework as the one we use. In fact, Lang’s lecture notes [37] served in many instances as a guideline for our work. However, ROW methods have also been applied to the Navier-Stokes equations by the different technique of first discretizing in space and then applying the ROW method to the resulting system of classic ordinary differential equations — the papers [49, 50] by Rang as well as the paper [32] by John, Matthies and Rang are good examples for that approach.

A common argument for avoiding W-methods with approximate Fréchet derivatives is that these methods lead to a loss of accuracy and/or stability — see again [37]. This is certainly true but we think that the potential numerical speedup which could be gained by, for instance, only very sporadically assembling the stiffness matrix is worth the effort of further studying these methods. Moreover, in our numerical experiments with several parabolic problems, the tested W-methods that only use approximations to the exact Fréchet derivatives do indeed produce promising results.

Besides providing a strong mathematical foundation for the semi-discretization in time of semilinear parabolic equations by ROW methods and W-methods, we also demonstrate that these methods are feasible in practical applications. This means, of course, that we require a spatial discretization as well, which could in turn influence the time-stepping procedure. It is, however, not within the scope of this work to provide an exact analysis of corresponding fully-discrete schemes.

Furthermore, since our focus lies on the temporal discretization, we carry out experiments in which the spatial error is kept small enough so that it can largely be neglected. In his lecture notes [37], Jens Lang points out that keeping the spatial error below some tolerance leads to a time integration procedure which is very similar to the semi-discrete one. He also shows in detail how for ROW methods, the purely temporal error estimates can be extended to fully-discrete estimates — with decoupled spatial and temporal error — if the spatial discretization fulfills certain requirements. Though we did not work out the details, it seems that his procedure is also feasible for W-methods.

We study numerical experiments with medium order ROW methods and W-methods that are applied to convection-diffusion equations, reaction-diffusion equations and the incompressible Navier-Stokes equations. The finite element toolkit Gascoigne 3D (see [20]) is used for implementation and — as mentioned above — the spatial error is kept small enough to be negligible and allows the focus to be on the temporal error. The experimental results mostly confirm our theoretical estimates and a comparison with the popular Crank-Nicholson time-stepping scheme suggests that the ROW methods and W-methods are feasible for the temporal discretization of these equations but also have certain limitations: in particular, W-methods do display serious stability issues in some applications — depending on how the exact Fréchet derivative is approximated. Thus, very small time steps might be required in some cases.

We now give a short overview of the contents: In **chapter 2** we present the natural-scientific background and motivation for our mathematical studies. Starting from a very general continuity equation, we briefly show how to derive the convection-diffusion and Navier-Stokes equations that provide the basic mathematical modeling of mass or heat transfer and incompressible flow.

As opposed to the previous, rather elementary formulation, the following **chapter 3** develops a more abstract framework and then demonstrates that the aforementioned physical applications are covered by this approach. After a short section on basic notation, we give a comprehensive

introduction to sectorial operators and show in detail that our convection-diffusion and the Stokes operator are of this type. To obtain a general and elegant notion of regularity in our abstract setting, we define fractional powers of sectorial operators with positive spectrum and explore how for the convection-diffusion operator and the Stokes operator the graph norms of these fractional powers relate to Sobolev norms.

We then provide a simple way to extend a certain class of unbounded sectorial operators to bounded operators on Gelfand triples. We also verify that the aforementioned operators are of this class, meaning that the construction — which we based on ideas from the paper [42] by Lubich and Ostermann — can be applied to them.

Later in this chapter, we use the previous definitions and results to give a general problem formulation for semilinear parabolic equations and list some commonly known existence, uniqueness and regularity results. Finally, we examine the incompressible Navier-Stokes equations in regard to this formulation. Based on example 3.8 in [27], we validate that it does indeed fit into our framework. We also show that our Oseen operator, which naturally emerges in linearizations of the Navier-Stokes equations, is a “well-behaved” sectorial operator in the sense that despite losing the symmetry, it keeps important properties of the Stokes operator pertaining to fractional powers.

Chapter 4 is entirely devoted to the description and examination of Rosenbrock-type methods and their application to ordinary differential equations as well as semilinear parabolic equations. In the beginning of the chapter, Rothe’s method (see [57] for the original paper) of first discretizing in time and then in space is explained. For the next three sections, we undertake an excursion into the field of ordinary differential equations. This is justified and useful because — as explained above — we formulate certain parabolic PDEs as SPEs, which can be seen as ODEs with values in Hilbert spaces.

We convey the idea behind Rosenbrock-type methods in two different ways which both boil down to working the Jacobian given by a possibly nonlinear ODE into the time-stepping scheme — for W-methods, the Jacobian can even be approximated. Next, we recall classic properties of ODE solvers, such as convergence order and stability, and give order conditions for ROW methods and W-methods. When examining these properties, we always keep in mind our eventual aim to treat parabolic PDEs — especially interesting in this context are the different notions of stiffness that we discuss in this and the following ODE section. That final ODE section focuses completely on stiff ODEs and how to define properties of ODE solvers that are more suited for these equations. The section can also be seen as a transition to the “infinitely stiff” parabolic PDEs — for example a spatially discretized heat equation is studied as a prime example of stiff ODEs.

In the last part of chapter 4, we finally leave the field of ODEs and apply ROW methods and W-methods to semilinear parabolic equations and give error estimates for the resulting semi-discretization. These estimates are from the paper [41] by Lubich and Ostermann. In a rather long section, we show in detail that the convection-diffusion equation, the incompressible Navier-Stokes equations and some types of reaction-diffusion equations meet the requirements posed in their paper. For W-methods, the choice of operators with which the exact Fréchet derivatives are approximated is, of course, extremely important — we inspect some feasible options and thoroughly prove that they fulfill the conditions posed in the paper [41].

We also discuss the problem of needing a very smooth exact solution in order to show these error estimates and explain what this requirement means in applications.

In order to obtain a concrete numerical solution, the stationary problems that result from the semi-discretization in time need to be discretized in space as well. In **chapter 5**, we give a short introduction to our method for doing that: the popular finite element method (FEM). Since the focus of this work lies on the temporal discretization, we do not delve into the details here. We only present the basic ideas and also restrict ourselves to those implementations of the finite element method that we need for our applications.

We begin the chapter by showing how the previously examined temporal discretization leads to successive stationary problems to be solved and how in the finite element method, approximate solutions to those stationary problems are sought in finite dimensional spaces that are built from a subdivision of the spatial domain into polytopes.

The next section of that chapter is largely based on the lecture notes [8] by Malte Braack. In that section, we first describe the type of mesh with which we cover a spatial domain and then we show how to construct piecewise polynomials, which are defined on such a mesh and approximate the exact solutions of the stationary problems. This process of finding approximate solutions is examined for both the stationary convection-diffusion equation as well as the Oseen-type equation, which arises from a linearization of the incompressible Navier-Stokes equations.

In the subsection on the Oseen-type equation we also introduce the local projection stabilization (LPS) technique (see [6] for the initial publication by Becker and Braack), which can be used to enable so-called “equal-order” finite element methods. At the end of the chapter, we then use ideas and results from Lang’s lecture notes [37] to take a short look at how fully-discrete error estimates could be constructed.

In **chapter 6** we study the performance of a few ROW methods and W-methods in several numerical experiments with parabolic PDEs. Most of the problems we examine in that chapter are designed to have a specific known solution — with the initial and boundary conditions as well as the right-hand side set accordingly. We design and numerically test three problems in that way: a convection-diffusion problem, a reaction-diffusion problem with zero-order nonlinearity and an incompressible Navier-Stokes problem.

All the Rosenbrock-type time-stepping schemes we test can be used both as ROW methods and as W-methods. Therefore, a main focus is to investigate the impact of only approximating the exact Fréchet derivatives by comparing, for a single method, the numerical performance with exact (i.e., using the method as ROW method) and inexact Fréchet derivatives (i.e. using the method as as W-method). Depending on the type of problem, our W-method variants range from occasionally building the stiffness matrix that corresponds to the exact Fréchet derivative to building only one stiffness matrix for the whole time-stepping procedure to dropping the convective term from the stiffness matrix entirely.

Most of the results from our experiments with known solutions suggest that using inexact Fréchet derivatives can lead to stability problems when the examined problem is convection-dominated and the convective term in the stiffness matrix is only approximated or omitted. We observe that for our convection-dominated problems, W-methods with inexact Fréchet derivatives do not produce feasible solutions for large time steps. However, the results show that for small and some

medium time steps, those W-methods produce numerical solutions with surprisingly small errors.

At the end of chapter 6, we display the results from numerical experiments with the well-known benchmark problem of a two-dimensional flow around a circular obstacle and the von Kármán vortex street forming behind the obstacle. Similarly to the experiments with known solutions and higher Reynolds number, our results here show that W-methods with inexact Fréchet derivatives may require relatively small time steps to produce feasible solutions. However, they also show that the larger the update frequency of the stiffness matrix, the larger the admissible time step size. In the benchmark problem we measure accuracy by comparing drag, lift and pressure difference against a reference solution that was calculated in a similar but slightly different setting. All tested methods — including the W-methods with inexact Fréchet derivatives and small enough time steps so that a feasible solution is produced — display a promisingly high accuracy.

Finally, we summarize the content and results of our work in **chapter 7** and offer some ideas about possible extensions and improvements.

2 Physical Background

In this chapter we want to illuminate the physical roots of the convection-diffusion and the incompressible Navier-Stokes equations and, in doing so, motivate the search for associated efficient numerical solvers.

The origin of these partial differential equations lies in an area of physics called *continuum mechanics*. In this field, matter is treated as continuous and not as the conglomerate of discrete particles that it really is. This approach has proven to be successful and accurate for describing many large scale physical phenomena. On microscopic scales, one would have to use concepts from other areas of physics such as *statistical mechanics* or *quantum mechanics*.

Our presentation is based on the beginning of chapter 14 in [8], on section 1.2 in [36] and on the explanation in [71]. As a warning to the reader we want to mention that this chapter is meant merely as an introductory demonstration of some principles and results from the field of continuum mechanics. Therefore, we will omit many mathematical and physical details. In particular, we always assume that the functions we use are well-defined and sufficiently differentiable. For more details on the derivation of the convection-diffusion and the Navier-Stokes equations, we refer to the above mentioned literature.

Some of the most fundamental equations in continuum mechanics are the so-called continuity equations. A general version that we use as a starting point can be written as follows:

$$\frac{\partial \phi}{\partial t} + \operatorname{div} f = s. \quad (2.1)$$

Here, $\phi = \phi(t, x)$ with $t > 0$ and $x \in \mathbb{R}^d$, $d \in \{2, 3\}$, is the (either scalar or vector-valued) density of some physical quantity, $f = f(t, x)$ is the (vector or matrix-valued) flux of that quantity and $s = s(t, x)$ are (either scalar or vector-valued) net sources and sinks of that quantity.

Roughly speaking, this equation states that a flowing or moving physical quantity — such as mass, energy, electric charge, momentum — has a rate of change within any given volume that is completely determined by inflow and outflow through the surface of that volume (described by the flux f) and any creation or removal of the physical quantity within that volume (described by the source term s). This is best seen in the following *integral form* of (2.1). Indeed, under the usual circumstances, the divergence theorem ensures that (2.1) is equivalent to

$$\frac{d}{dt} \left(\int_V \phi \, dV \right) + \int_{\partial V} (f \cdot n) \, d(\partial V) = \int_V s \, dV$$

holding for all $V \subset \mathbb{R}^d$ that are “sufficiently nice” and in particular have a boundary ∂V with an outward pointing unit normal field n . A single one of these V is often called a *control volume*. For illustration see figure 2.1 below, a slightly altered copy of a picture from page 4 of [36].

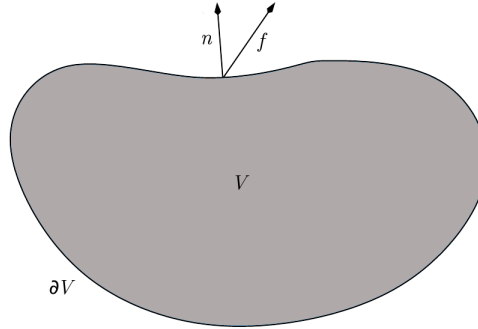


Figure 2.1: A fixed control volume V with boundary ∂V and outward pointing unit normal field n . f denotes the flux.

In the next sections, we will use (2.1) to derive the two main equations that interest us.

2.1 Mass and Heat transfer

Here we study the case that the density ϕ in (2.1) is given as $\phi = u = \rho\gamma$, where γ is the concentration (amount per unit mass) of a scalar quantity — such as heat or some chemical species — and ρ is the mass density (unit mass per unit volume) of the material that contains and carries that quantity. The hypothesis is now that the flux f in (2.1) has both a convective and a diffusive component. More specifically,

$$f = f_C + f_D = bu - \mathcal{D}\rho\nabla\gamma, \quad (2.2)$$

where we assume that the carrying material has a known velocity b which “convects” our quantity of interest and also a possibly time and space dependent diffusion coefficient \mathcal{D} that describes the diffusive processes. The specific form of the term f_D is known as Fick’s law in the case of mass diffusion and as Fourier’s law in the case of heat conduction.

Inserting $\phi = u = \rho\gamma$ and (2.2) into (2.1) yields

$$\frac{\partial(\rho\gamma)}{\partial t} + \operatorname{div}(b\rho\gamma) - \operatorname{div}(\mathcal{D}\rho\nabla\gamma) = s.$$

Assuming that the mass density ρ is constant and the velocity field b is incompressible (i.e., $\operatorname{div} b = 0$), then

$$\frac{\partial u}{\partial t} + (b \cdot \nabla)u - \operatorname{div}(\mathcal{D}\nabla u) = s. \quad (2.3)$$

Even though we used $\operatorname{div} b = 0$ for the derivation of this equation, we will often examine the interesting case of allowing $\operatorname{div} b \neq 0$ in (2.3) as well.

Some reordering of the terms and further simplifying the equation by assuming $\mathcal{D} \equiv \nu > 0$

as well as $s(t, x) = -c(x)u(t, x) + f(t)$ for some scalar functions c (zero-order coefficient) and f (outside force) finally provides us with the convection-diffusion equation that we use as basis for our more detailed mathematical analysis:

$$\frac{\partial u}{\partial t} - \nu \Delta u + (b \cdot \nabla)u + cu = f. \quad (2.4)$$

PDEs like this one are often studied with the time t varying in some bounded interval $[0, T]$ ($T > 0$ being the final time) and the spatial coordinate x varying in a bounded domain $\Omega \subset \mathbb{R}^d$ with “sufficiently smooth” boundary. In order to create a well-posed problem however, we then have to also provide an initial condition (i.e., we need to specify $u(t = 0)$) and a suitable boundary condition, such as a Dirichlet boundary condition (u is specified on $\partial\Omega$), a Neumann boundary condition ($\frac{\partial u}{\partial n}$ is specified on $\partial\Omega$) or a combination of the two.

In sections 3.3 and 3.7 we will show that under certain requirements on the coefficients and the domain, (2.4) can easily be treated as an ordinary differential equation with values in an infinite dimensional Hilbert space.

2.2 Incompressible Flow

In this section we first choose the flux f in (2.1) to be $f = \phi v$, where ϕ is the (vector or scalar-valued) density of some physical quantity and v is a velocity field. Thus, we obtain the equation

$$\frac{\partial \phi}{\partial t} + \operatorname{div}(\phi v) = s, \quad (2.5)$$

from which we will then derive the incompressible Navier-Stokes equations after employing different choices for the density ϕ and the source term s and making several further assumptions. Notice that if ϕ is vector-valued, ϕv has to be understood as a dyadic product — with the divergence of the resulting second rank tensor being a first rank tensor, i.e., a vector.

In a first and easy step, we account for *conservation of mass* by setting $s = 0$ and substituting $\phi = \rho$ in (2.5), with ρ being the (scalar-valued) mass density. This leads to

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho v) = 0. \quad (2.6)$$

If one assumes that the density ρ is constant when following the path of any fluid element, i.e., the so-called *material derivative* $\frac{D}{Dt}$ of ρ fulfills

$$\frac{D\rho}{Dt} := \frac{\partial \rho}{\partial t} + (v \cdot \nabla)\rho = 0, \quad (2.7)$$

one then obtains the incompressibility condition in its well-known form:

$$\operatorname{div} v = 0. \quad (2.8)$$

We note that (2.7) is fulfilled in particular if ρ is constant in time *and* space.

Now we turn our attention to the *conservation of momentum*. Here we set $\phi = \rho v$ in (2.5) and use definitions and identities from vector calculus as well as the conservation of mass as given in (2.8) to acquire

$$\rho \left(\frac{\partial v}{\partial t} + (v \cdot \nabla) v \right) = s, \quad (2.9)$$

or equivalently, using the notion of the material derivative,

$$\rho \frac{Dv}{Dt} = s. \quad (2.10)$$

Using basic principles from continuum mechanics, it can be shown that the source term s in (2.10) has to have the following form:

$$s = \operatorname{div} \sigma + f. \quad (2.11)$$

The matrix-valued term σ is called the Cauchy stress tensor and describes surface forces, while the term f describes body forces, such as gravity and electromagnetic forces. Piecing together (2.9) and (2.11) does not yet produce a Navier-Stokes momentum equation.

We still need to specify the form of σ . There are several different possibilities to do that — leading to different equations such as Newtonian and non-Newtonian Navier-Stokes equations or the Euler equation. In this introduction, we restrict ourselves to the choice for σ that leads to the standard incompressible Navier-Stokes equations for Newtonian fluids. Those equations are obtained when assuming that

$$\sigma = -pI_d + \mu \nabla v, \quad (2.12)$$

with the (scalar-valued) pressure p and a constant viscosity $\mu > 0$. Renaming $p \rightarrow \frac{p}{\rho}$ and $f \rightarrow \frac{f}{\rho}$ as well as defining the kinematic viscosity $\nu := \frac{\mu}{\rho}$, we finally assemble the familiar incompressible Navier-Stokes equations from (2.9), (2.11), (2.12) and (2.8):

$$\frac{\partial v}{\partial t} + (v \cdot \nabla) v - \nu \Delta v + \nabla p = f, \quad (2.13)$$

$$\operatorname{div} v = 0. \quad (2.14)$$

As in section 2.1, the equations (2.13) and (2.14) need to be equipped with an initial condition and (for bounded domains) a boundary condition in order to form a well-posed problem.

In the following chapter, we will examine some mathematical properties of the convection-diffusion and the incompressible Navier-Stokes equations in a more rigorous way. Specifically, we want to show how these equations can be viewed, in a sense, as “infinitely stiff” ordinary differential equations and can consequently be treated as such when discretizing them in time.

3 Semilinear Parabolic Model

We now take a step back from the concrete physical problems that we looked at in the previous chapter and introduce the abstract mathematical framework of semilinear parabolic equations. During the course of this chapter it will become apparent that the equations from the previous chapter do indeed fit into that abstract framework and that within that framework we can handle partial differential equations very similarly to *ordinary* differential equations. This will be very helpful for our studies because we are primarily interested in the dynamic processes modeled by these equations and, correspondingly, their *temporal* discretization.

Furthermore, a significant amount of literature on the discretization in time of parabolic partial differential equations uses this framework. In particular, our goal is to rigorously prove that we may apply the semi-discrete (in time) error bounds that are given in the paper [41] by Lubich and Ostermann. Their results are formulated for a certain class of these semilinear parabolic equations, and we will show in detail that the equations introduced in chapter 2 are in that class.

Before we delve into the theory of sectorial operators and semilinear equations, we will recall some notations, definitions and results from functional analysis.

3.1 Basic Notations and Mathematical Concepts

In this section we want to establish our notation and basic definitions. Furthermore, we are going to state some results from operator theory and functional analysis that we will often refer to.

For the sake of brevity, however, we do not list all definitions and theorems that we use — i.e., we omit things such as L^p -spaces, the Hölder inequality, the Hahn-Banach theorem, the Riesz representation theorem, the Lax-Milgram theorem and others. When using these, we simply trust in the reader's familiarity with the subject. Similarly, when we *do* include a well-known result, we often omit the proof. For more details, we refer to one of the many textbooks or lecture notes on these topics, such as [35] for operator theory, [3] for functional analysis, [16] for partial differential equations, [60] or [73] for evolution equations and [1] for Sobolev spaces.

Definition 3.1.1 (Densely-Defined, Closed and Bounded Linear Operators). *Let X and Y be normed spaces over \mathbb{K} , $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, and let $\mathbb{A} : D(\mathbb{A}) \subset X \rightarrow Y$ be a linear operator with the domain $D(\mathbb{A})$ of \mathbb{A} being a subspace of X . Then we call \mathbb{A}*

- (i) *densely-defined if $D(\mathbb{A})$ is dense in X , i.e., if $\overline{D(\mathbb{A})}^{\|\cdot\|_X} = X$,*
- (ii) *closed if its graph is closed, i.e., if for all $(x_n)_{n \in \mathbb{N}} \in (D(\mathbb{A}))^{\mathbb{N}}$ with $\lim_{n \rightarrow \infty} \|x_n - x\|_X = 0$ for*

some $x \in X$ and $\lim_{n \rightarrow \infty} \|\mathbb{A}x_n - y\|_Y = 0$ for some $y \in Y$, we have $x \in D(\mathbb{A})$ and $Tx = y$,

(iii) bounded if $D(\mathbb{A}) = X$ and

$$\underbrace{\sup_{x \in X \setminus \{0\}} \frac{\|\mathbb{A}x\|_Y}{\|x\|_X}}_{=: \|\mathbb{A}\|_{\mathcal{L}(X, Y)}} < \infty,$$

with $\mathcal{L}(X, Y)$ denoting the normed space over \mathbb{K} of all linear operators mapping from X to Y that are bounded in the above defined sense.

Notice that bounded linear operators are both densely-defined and closed.

We now turn our attention to Hilbert spaces — usually we always work with those instead of just Banach spaces.

Remark 3.1.1. *Normally, the Hilbert spaces we work with are — unless otherwise stated — vector spaces over the complex numbers. Therefore, we have to choose which component of the accompanying dot product is linear and which is antilinear. To reduce clutter in our presentation, we hereby state that **throughout this work, we assume that any dot products are linear in the first component and antilinear in the second.***

Next, we give a definition and some facts about the adjoint of a linear operator between Hilbert spaces.

Definition 3.1.2 (Adjoint Operator in a Hilbert Space Setting). *Let H_1, H_2 be Hilbert spaces and let $\mathbb{A} : D(\mathbb{A}) \subset H_1 \rightarrow H_2$ be a linear and densely-defined operator. We now define the adjoint $\mathbb{A}^* : D(\mathbb{A}^*) \subset H_2 \rightarrow H_1$ of \mathbb{A} by setting*

$$D(\mathbb{A}^*) := \{y \in H_2 : \exists! x_y \in H_1 \forall x \in D(\mathbb{A}) : (\mathbb{A}x, y)_{H_2} = (x, \underbrace{x_y}_{=: \mathbb{A}^* y})_{H_1}\}.$$

Remark 3.1.2. *Let H_1, H_2 be Hilbert spaces and let $\mathbb{A} : D(\mathbb{A}) \subset H_1 \rightarrow H_2$ be a linear and densely-defined operator. Then we have*

(i)

$$\begin{aligned} D(\mathbb{A}^*) &= \{y \in H_2 : \exists x_y \in H_1 \forall x \in D(\mathbb{A}) : (\mathbb{A}x, y)_{H_2} = (x, x_y)_{H_1}\} \\ &= \{y \in H_2 : \exists C_y > 0 \forall x \in D(\mathbb{A}) : |(\mathbb{A}x, y)_{H_2}| \leq C_y \|x\|_{H_1}\}. \end{aligned}$$

(ii) \mathbb{A}^* is closed.

Proof. *The first part is a straight consequence of the Riesz representation theorem and the density of $D(\mathbb{A})$ in H_1 .*

For the second part let $(y_n)_{n \in \mathbb{N}} \in (D(\mathbb{A}^*))^{\mathbb{N}}$ with $\lim_{n \rightarrow \infty} \|y_n - y\|_{H_2} = 0$ for some $y \in H_2$ and let $\lim_{n \rightarrow \infty} \|\mathbb{A}^* y_n - x\|_{H_1} = 0$ for some $x \in H_1$. It is then easy to see that for all $\hat{x} \in D(\mathbb{A})$ we have

$$(\mathbb{A}\hat{x}, y)_{H_2} = (\hat{x}, x)_{H_1}.$$

Using the first part of this remark and the definition of the adjoint, we thus obtain $y \in D(\mathbb{A}^*)$ and $\mathbb{A}^* y = x$.

The concept of a self-adjoint operator is not as straightforward for unbounded operators, as it is for bounded ones. This is because we always have to pay attention to the domain as well here — leading to the notion of merely *symmetric* operators.

Definition 3.1.3 (Symmetric and Self-Adjoint Linear Operators). *Let H be a Hilbert space and let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a linear and densely-defined operator. We then call \mathbb{A} symmetric if for all $x, y \in D(\mathbb{A})$ we have*

$$(\mathbb{A}x, y)_H = (x, \mathbb{A}y)_H.$$

We call \mathbb{A} self-adjoint if it is symmetric and we have $D(\mathbb{A}^) = D(\mathbb{A})$, i.e., if we have the equality $\mathbb{A} = \mathbb{A}^*$ with equal domains.*

For examining evolution equations, the notion of a *triplet of spaces*, which we introduce next, is particularly useful. We start by clarifying our notations of (anti-)dual spaces and the Riesz isomorphism.

Definition 3.1.4 (Dual and Antidual Space). *Let X be a normed space over \mathbb{K} with $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$.*

- *Then we call the normed space $\mathcal{L}(X, \mathbb{K})$ the dual space of X as usual.*
- *If $\mathbb{K} = \mathbb{C}$, we denote by X' the antidual space of X , i.e., the space of all continuous antilinear functions mapping from X to \mathbb{C} . Naturally, we equip X' with the same norm as the dual space, i.e., with the $\|\cdot\|_{\mathcal{L}(X, \mathbb{C})}$ norm.*
- *If X is a normed space over the real numbers, we will also use the notation $X' = \mathcal{L}(X, \mathbb{R})$.*

Definition 3.1.5 (Riesz Isomorphism). *Let H be a Hilbert space and let $\Phi_H : H \rightarrow H'$ be defined through*

$$\Phi_H(h) := (h, \cdot)_H \quad \text{for all } h \in H.$$

Then we call Φ_H the Riesz isomorphism for H .

Notice that for complex Hilbert spaces, other authors often define a Riesz *anti*isomorphism into the dual space, as opposed to the above defined isomorphism into the antidual space. We opted for an isomorphism instead, letting us properly identify a Hilbert space with its antidual via that isomorphism.

A triplet of spaces is now defined as follows:

Definition 3.1.6 (Triplet of Spaces). *Let V, H be Hilbert spaces. Then we call (V, H, V') a triplet of spaces if*

- (i) *V is a subspace of H with $\overline{V}^{\|\cdot\|_H} = H$, i.e., if V is dense in H ,*
- (ii) *there is a constant $C > 0$ so that $\|x\|_H \leq C\|x\|_V$, i.e., if the $\|\cdot\|_H$ -norm is bounded by the $\|\cdot\|_V$ -norm.*

In the literature, triplets with the above properties are frequently called Gelfand triples. We use a different name here because in the common definition of Gelfand triples, the subspace V may just be a locally convex topological vector space — a generalization that is not needed for this work and would only require a more complicated definition of the (anti-)dual space.

The following lemma shows why including the (anti-)dual space in the definition and using the name *triplet* of spaces makes sense.

Lemma 3.1.1. *Let (V, H, V') be a triplet of spaces with $C > 0$ denoting the constant from part (ii) of the above definition. Then the mapping $\iota : H' \rightarrow V', h' \mapsto \iota(h') := h'|_V$ is a well-defined linear and continuous injection with*

$$\overline{\iota(H')}^{\|\cdot\|_{V'}} = V',$$

$$\|\iota(h')\|_{V'} \leq C\|h'\|_{H'} \quad \text{for all } h' \in H' \quad (3.1)$$

and

$$\langle \iota(\Phi_H(h)), v \rangle = (h, v)_H \quad \text{for all } h \in H \text{ and } v \in V.$$

This can be proven from the definitions in a very straightforward, albeit technical way.

Now, if (V, H, V') is a triplet of spaces and we identify H with its (anti-)dual space via the Riesz isomorphism Φ_H as usual, we may use the above result to write

$$V \xhookrightarrow{d} H \xhookrightarrow{d} V'$$

where \xhookrightarrow{d} denotes a dense and continuous embedding. In particular, if $C > 0$ denotes the constant from definition 3.1.6 and ι the mapping from lemma 3.1.1, then that lemma provides us for any $v \in V$ with the bound

$$\frac{1}{C} \|\iota(\Phi_H(v))\|_{V'} \leq \|v\|_H \leq C\|v\|_V.$$

Hence, we will usually work with (V, H, V') as if we have the dense inclusions $V \subset H \subset V'$ and the bounds

$$\frac{1}{C} \|v\|_{V'} \leq \|v\|_H \leq C\|v\|_V \quad (3.2)$$

for all $v \in V$. This means that we opt to identify V with its image under the mapping $\iota \circ \Phi_H$. We want to stress once again that one has to be mindful of the underlying identifications and

embeddings. In particular, one normally has $(\iota \circ \Phi_H)|_V \neq \Phi_V$, with Φ_V being the Riesz isomorphism for V . So we do **not** identify V with its (anti-)dual space V' via Φ_V .

We now briefly introduce a generalization of the exponential function which lets us represent solutions to many different evolution equations.

Definition 3.1.7 (Analytic Semigroup and its Generator). *Let X be a Banach space, let $\psi \in (0, \frac{\pi}{2})$ and let $M := \{\lambda \in \mathbb{C} : |\arg(\lambda)| < \psi\} \cup \{0\}$. Furthermore, let $T(z) \in \mathcal{L}(X, X)$ for each $z \in M$. Then we call the family $(T(z))_{z \in M}$ an analytic semigroup on X if*

$$(i) \quad T(0) = Id_X \text{ and } T(z_1)T(z_2) = T(z_1 + z_2) \text{ for all } z_1, z_2 \in M,$$

$$(ii) \quad \text{the mapping } M \setminus \{0\} \ni z \mapsto T(z) \in \mathcal{L}(X, X) \text{ is analytic,}$$

$$(iii) \quad \text{for any } \psi' \in (0, \psi) \text{ and all } x \in X \text{ we have}$$

$$\lim_{\substack{z \rightarrow 0 \\ z \in M(\psi')}} \|T(z)x - x\|_X = 0,$$

$$\text{where } M(\psi') := \{\lambda \in \mathbb{C} : |\arg(\lambda)| \leq \psi'\} \setminus \{0\}.$$

We say that an operator $\mathbb{A} : D(\mathbb{A}) \subset X \rightarrow X$ is the generator of the analytic semigroup $(T(z))_{z \in M}$ if

$$D(\mathbb{A}) = \{x \in X : \exists y \in X : \lim_{h \rightarrow 0^+} \|h^{-1}(T(h)x - x) - \underbrace{y}_{=\mathbb{A}x}\|_X = 0\}.$$

We want to mention that there are more general concepts of this, such as strongly continuous semigroups. We will see in section 3.2, though, that analytic semigroups fit exactly into our mathematical framework.

The final part of this section will be on the well-known Sobolev spaces. For the sake of clarity and completeness, we go over the basic definitions, results and notations. Before we state the definition of Sobolev spaces, however, we list our notations for the different types of derivatives that we need throughout this work.

Remark 3.1.3.

- Let V, W be normed vector spaces and let $U \subset V$ be open. Then, if it exists, we denote the first, second, third, ... Fréchet derivative of a function $u : U \rightarrow W$ by

$$D^{(1)}u, D^{(2)}u, D^{(3)}u, \dots$$

- Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, let $U \subset \mathbb{K}$ be open and let W be a normed vector space. Then, if it exists, we also denote the first, second, third, ... Fréchet derivative of a function $U \ni x \mapsto u(x) \in W$ by

$$\frac{du}{dx}, \frac{d^2u}{dx^2}, \frac{d^3u}{dx^3}, \dots$$

Note that the notation is a bit vague in that it depends on the variable which is used in the definition of u . That is, if instead of x , other variables such as z (often used if $\mathbb{K} = \mathbb{C}$) or t (often used to symbolize time) are used, then these derivatives are normally denoted as $\frac{du}{dz}$ or $\frac{du}{dt}$ correspondingly.

• Let $d \in \mathbb{N}$, $j, k, l \in \{1, \dots, d\}$, $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$ and let $\Omega \subset \mathbb{R}^d$ be open. Then, if they exist, we denote the partial derivatives and the higher and mixed partial derivatives of a function $u : \Omega \rightarrow \mathbb{K}$ in the usual manner, i.e., by

$$\frac{\partial u}{\partial x_j}, \frac{\partial^2 u}{\partial x_j \partial x_k}, \frac{\partial^3 u}{\partial x_j^2 \partial x_k}, \frac{\partial^3 u}{\partial x_j \partial x_k \partial x_l}, \dots$$

For $n \in \mathbb{N}_0$ and a multi-index $\alpha \in \mathbb{N}_0^n$ we also employ the standard notation

$$D^\alpha u = \frac{\partial^{|\alpha|} u}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}.$$

Moreover, we use the standard symbols from vector calculus, i.e., ∇u for the gradient of u , as well as ∇v for the Jacobian and $\operatorname{div} v = \nabla \cdot v$ for the divergence of a vector-valued function $v : \Omega \rightarrow \mathbb{K}^r$, $r \in \mathbb{N}$. We further use the notations Δu , Δv and $(b \cdot \nabla)u$, $(b \cdot \nabla)v$ for some $b \in \mathbb{K}^r$ with their respective meanings for scalar and vector-valued functions in the usual way.

• We will also use the aforementioned partial derivative notation on functions for which these partial derivatives do not exist in the classical, but merely in a weak sense. In that case, we, of course, mean the weak derivative.

If $T \in \mathbb{R}_{>0}$ and X is a Banach space, then we denote the weak derivatives of a function $u : [0, T] \rightarrow X$, which are defined in definition 3.1.10 below, by

$$u = u^{(0)}, u' = u^{(1)}, u'' = u^{(2)}, u''' = u^{(3)}, u^{(4)}, \dots$$

In the following definition of the Sobolev spaces, we mention (without proof) some not so trivial but very well-known facts about these spaces — namely that they are complete and that most of them are separable.

Definition 3.1.8 (Sobolev Spaces). Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, $d \in \mathbb{N}$ and let $\Omega \subset \mathbb{R}^d$ be open. Then we define for all $m \in \mathbb{N}_0$ and $p \in \mathbb{N} \cup \{\infty\}$ the space

$$W^{m,p}(\Omega, \mathbb{K}) := \{u \in L^p(\Omega, \mathbb{K}) : D^\alpha u \in L^p(\Omega, \mathbb{K}) \text{ for all } \alpha \in \mathbb{N}_0^d \text{ with } \|\alpha\|_1 \leq m\}.$$

and equip it with the norm

$$W^{m,p}(\Omega, \mathbb{K}) \ni u \mapsto \|u\|_{W^{m,p}(\Omega, \mathbb{K})} := \begin{cases} \left(\sum_{\substack{\alpha \in \mathbb{N}_0^d \\ \|\alpha\|_1 \leq m}} \|D^\alpha u\|_{L^p(\Omega, \mathbb{K})}^p \right)^{\frac{1}{p}} & \text{if } p < \infty \\ \max_{\substack{\alpha \in \mathbb{N}_0^d \\ \|\alpha\|_1 \leq m}} \|D^\alpha u\|_{L^\infty(\Omega, \mathbb{K})} & \text{if } p = \infty \end{cases}$$

to make it a Banach space that is separable if $p < \infty$.

Furthermore, we write

$$H^m(\Omega, \mathbb{K}) := W^{m,2}(\Omega, \mathbb{K})$$

and note that the $\|\cdot\|_{H^m(\Omega, \mathbb{K})}^p = \|\cdot\|_{W^{m,2}(\Omega, \mathbb{K})}^p$ norm is induced by the dot product

$$(u, v)_{H^m(\Omega, \mathbb{K})} = \sum_{\substack{\alpha \in \mathbb{N}_0^d \\ \|\alpha\|_1 \leq m}} \int_{\Omega} D^{\alpha} u \overline{D^{\alpha} v} d\Omega \quad \text{for all } u, v \in H^m(\Omega, \mathbb{K}),$$

so that $H^m(\Omega, \mathbb{K})$ equipped with this dot product is a separable Hilbert space.

To make sense of boundary values, we need the following well-known result (for a proof see for example [1]):

Theorem 3.1.1. *Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, $d \in \mathbb{N}$ and let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then there exists a bounded linear operator $\gamma : H^1(\Omega, \mathbb{K}) \rightarrow L^2(\partial\Omega, \mathbb{K})$ — a so-called trace operator — with*

$$\gamma(u) = u|_{\partial\Omega}$$

for all $u \in H^1(\Omega, \mathbb{K}) \cap C(\overline{\Omega}, \mathbb{K})$.

This enables the definition:

Definition 3.1.9 (The Space $H_0^1(\Omega, \mathbb{K})$). *Let $\mathbb{K} \in \{\mathbb{R}, \mathbb{C}\}$, $d \in \mathbb{N}$ and let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Then we define the space*

$$H_0^1(\Omega, \mathbb{K}) = \{u \in H^1(\Omega, \mathbb{K}) : \gamma(u) = 0\}$$

and equip it with the $(\cdot, \cdot)_{H^1}$ dot product, making it a separable Hilbert space.

In order to shorten the presentation, we now introduce some abbreviating notation.

Remark 3.1.4. *Let $d \in \mathbb{N}$ and let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. Moreover, let $m \in \mathbb{N}_0$ and $p \in \mathbb{N} \cup \{\infty\}$. Unless otherwise specified, we then use the following abbreviations:*

$$\begin{aligned} L^p(\Omega) &:= L^p(\Omega, \mathbb{C}) & \text{with } \|\cdot\|_{L^p} &:= \|\cdot\|_{L^p(\Omega, \mathbb{C})} \text{ and } (\cdot, \cdot)_{L^2} := (\cdot, \cdot)_{L^2(\Omega, \mathbb{C})}, \\ W^{m,p}(\Omega) &:= W^{m,p}(\Omega, \mathbb{C}) & \text{with } \|\cdot\|_{W^{m,p}} &:= \|\cdot\|_{W^{m,p}(\Omega, \mathbb{C})}, \\ H^m(\Omega) &:= H^m(\Omega, \mathbb{C}) & \text{with } \|\cdot\|_{H^m} &:= \|\cdot\|_{H^m(\Omega, \mathbb{C})} \text{ and } (\cdot, \cdot)_{H^m} := (\cdot, \cdot)_{H^m(\Omega, \mathbb{C})}, \\ H_0^1(\Omega) &:= H_0^1(\Omega, \mathbb{C}), \\ H^{-1}(\Omega) &:= [H_0^1(\Omega)]' & \text{with } \|\cdot\|_{H^{-1}} &:= \|\cdot\|_{[H_0^1(\Omega)]'}. \end{aligned}$$

For any $r \in \mathbb{N}$ and functions $u = (u_1, \dots, u_r) \in W^{m,p}(\Omega)^r$, $v = (v_1, \dots, v_r) \in W^{m,p}(\Omega)^r$ we use the following norms and dot products on the product spaces:

$$\begin{aligned} \|u\|_{W^{m,p}(\Omega)^r}^2 &:= \sum_{j=1}^r \|u_j\|_{W^{m,p}}^2, \\ (u, v)_{H^m(\Omega)^r} &:= \sum_{j=1}^r (u_j, v_j)_{H^m}. \end{aligned}$$

In our notation we will often not differentiate between the norms/dot products for vector-valued and scalar-valued functions, i.e., we usually write $\|\cdot\|_{L^p}$ instead of $\|\cdot\|_{L^p(\Omega)^r}$, $\|\cdot\|_{W^{m,p}}$ instead of $\|\cdot\|_{W^{m,p}(\Omega)^r}$, $\|\cdot\|_{H^{-1}}$ instead of $\|\cdot\|_{H^{-1}(\Omega)^r}$, $(\cdot, \cdot)_{L^2}$ instead of $(\cdot, \cdot)_{L^2(\Omega)^r}$ and $(\cdot, \cdot)_{H^m}$ instead of $(\cdot, \cdot)_{H^m(\Omega)^r}$.

For adequately handling evolution equations we now define some so-called Bochner spaces. However, for the sake of brevity, we do not go into the details of the Bochner integral. For more information on that subject we refer the interested reader to textbooks such as [16] — see specifically sections E.5 and 5.9.2. of that book.

Definition 3.1.10 (Bochner Spaces). *Let X be a Banach space and let $T \in \mathbb{R}_{>0}$. We then define the spaces*

$$L^2(0, T; X) := \{u : [0, T] \rightarrow X : u \text{ is Bochner measurable and } \int_0^T \|u(t)\|_X^2 dt < \infty\},$$

$$C([0, T]; X) := \{u : [0, T] \rightarrow X : u \text{ is continuous}\}$$

and equip them with the norms

$$L^2(0, T; X) \ni u \mapsto \|u\|_{L^2(0, T; X)} := \left(\int_0^T \|u(t)\|_X^2 dt \right)^{\frac{1}{2}},$$

$$C([0, T]; X) \ni u \mapsto \|u\|_{C([0, T]; X)} := \max_{t \in [0, T]} \|u(t)\|_X.$$

We say that $u \in L^2(0, T; X)$ has the n -th weak derivative $u^{(n)} \in L^2(0, T; X)$ for some $n \in \mathbb{N}$ if for all $\phi \in C_0^\infty(0, T)$ we have

$$\int_0^T u(t) \frac{d^n \phi}{dt^n}(t) dt = (-1)^n \int_0^T u^{(n)}(t) \phi(t) dt.$$

For the first few weak derivatives we write $u^{(0)} := u$, $u' := u^{(1)}$, $u'' := u^{(2)}$, $u''' := u^{(3)}$ in the usual manner.

We now define for all $m \in \mathbb{N}_0$ the spaces

$$H^m(0, T; X) := \{u \in L^2(0, T; X) : u^{(j)} \in L^2([0, T]; X) \text{ for all } j \in \{0, \dots, m\}\},$$

$$C^m([0, T]; X) := \{u \in C([0, T]; X) : u^{(j)} \in C([0, T]; X) \text{ for all } j \in \{0, \dots, m\}\}$$

and equip them with the norms

$$H^m(0, T; X) \ni u \mapsto \|u\|_{H^m(0, T; X)} := \left(\sum_{j=0}^m \|u^{(j)}\|_{L^2(0, T; X)}^2 \right)^{\frac{1}{2}},$$

$$C^m([0, T]; X) \ni u \mapsto \|u\|_{C^m([0, T]; X)} := \sum_{j=0}^m \|u^{(j)}\|_{C([0, T]; X)}$$

to make them Banach spaces. It is easy to see that $H^m(0, T; X)$ is even a Hilbert space if X is one.

Next, we look at a few of the many available embedding and approximation results for Sobolev and Bochner spaces.

Theorem 3.1.2.

- (i) Let $d \in \mathbb{N}$, $m, j \in \mathbb{N}_0$, $p, q \in [1, \infty)$ and let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain. With \hookrightarrow denoting a continuous embedding we then have:

$$\begin{aligned} W^{m+j,p}(\Omega) &\hookrightarrow W^{m,q}(\Omega) && \text{if } jp \leq d \text{ and } p \leq q \leq \frac{dp}{d-jp}, \\ W^{m+j,p}(\Omega) &\hookrightarrow C^m(\Omega) && \text{if } jp > d. \end{aligned}$$

Furthermore, we have:

$$\begin{aligned} W^{m,p}(\Omega) &= \overline{C^\infty(\Omega) \cap W^{m,p}(\Omega)}^{\|\cdot\|_{W^{m,p}}}, \\ L^p(\Omega) &= \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{L^p}}, \\ H_0^1(\Omega) &= \overline{C_0^\infty(\Omega)}^{\|\cdot\|_{H^1}}. \end{aligned}$$

- (ii) Let $T \in \mathbb{R}_{>0}$, $m \in \mathbb{N}$ and let X be a Banach space. With \hookrightarrow again denoting a continuous embedding we then have:

$$H^m(0, T; X) \hookrightarrow C^{m-1}([0, T]; X).$$

A proof of (i) can be obtained for example from [63]. Part (ii) is theorem A.5 in [37].

In the next section we delve a bit deeper into the theory of analytic semigroups (introduced above in definition 3.1.7) and the so-called *sectorial operators*.

3.2 Sectorial Operators

Roughly speaking, a sectorial operator is a — possibly unbounded — linear operator which has properties (a specific localization of its spectrum and a specific bound for its resolvent) that make it possible to use a generalization of Cauchy's integral formula to define an exponential function, fractional powers and other functions of the operator by integrating along a contour around its spectrum. We now give a precise mathematical formulation of this idea by stating one definition of sectorial operators and some known results about these operators.

Note that the definition of sectorial operators often differs slightly from author to author. Since a substantial portion of the following sections on sectorial operators and fractional powers of those operators is based on the well-known lecture notes by Henry [27], we use his definition of sectorial operators.

Throughout this section, H will denote a separable complex Hilbert space. We begin by remembering an important definition from the theory of unbounded operators.

Definition 3.2.1 (Resolvent Set and Spectrum). *Let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a linear, closed and densely defined operator. The resolvent set of \mathbb{A} is defined as*

$$\rho(\mathbb{A}) := \{\lambda \in \mathbb{C} : \lambda - \mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H \text{ is a bijection and } (\lambda - \mathbb{A})^{-1} \in \mathcal{L}(H, H)\}.$$

The spectrum of \mathbb{A} is then simply defined as the complement:

$$\sigma(\mathbb{A}) := \mathbb{C} \setminus \rho(\mathbb{A}).$$

Furthermore, for any $\phi \in (0, \pi)$ and $a \in \mathbb{R}$ we use the following Notation.

Definition 3.2.2 (Sector). *Let $\phi \in (0, \pi)$, $a \in \mathbb{R}$. We then define*

$$S_{a,\phi} := \{\lambda \in \mathbb{C} : \phi \leq |\arg(\lambda - a)| \leq \pi \wedge \lambda \neq a\},$$

and call $S_{a,\phi}$ a sector.

This is a subset of the complex plane, symmetric about the real axis, with opening angle $2\pi - 2\phi$.

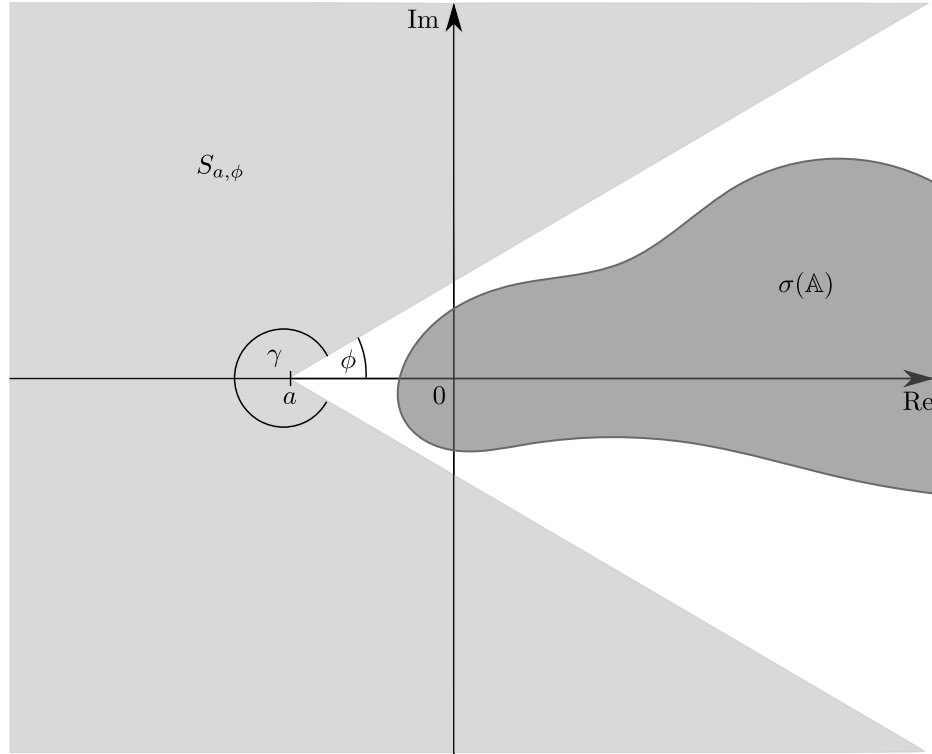


Figure 3.1: A sector $S_{a,\phi}$ with opening angle $\gamma := 2\pi - 2\phi$

Now we have everything we need to define sectorial operators:

Definition 3.2.3 (Sectorial Operator). *Let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a linear operator. We call \mathbb{A} a sectorial operator (on H) if*

- (i) $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ is closed and densely defined,
- (ii) there exist $\phi \in (0, \frac{\pi}{2})$ and $a \in \mathbb{R}$ so that $S_{a,\phi} \subset \rho(\mathbb{A})$,
- (iii) there exists $M_{\mathbb{A}} > 0$ so that for all $\lambda \in S_{a,\phi}$ the following resolvent bound holds:

$$\|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(H,H)} \leq \frac{M_{\mathbb{A}}}{|\lambda - a|}.$$

We will sometimes call $M_{\mathbb{A}}$ a sectoriality constant of \mathbb{A} .

Note that in the definition of sectorial operators, we required the opening angle $\gamma = 2\pi - 2\phi$ of the sector to be at least π . In the literature, sometimes other opening angles are permitted as well, and/or the sector is defined in a different way. Namely, the sector might be expected to **include** the spectrum as opposed to exclude it, as is the case in our definition. In all of our applications however, the restriction $\phi < \frac{\pi}{2}$ is not problematic — furthermore, it allows for an elegant correspondence between sectorial operators and the generators of analytic semigroups as demonstrated below.

A lot of theory is available on how to possibly define functions of sectorial operators (functional calculus). For a very detailed and careful approach we suggest looking into [24]. We are only interested in two main applications of functional calculus, though. We need:

- (a) the **exponential function** $e^{-t\mathbb{A}}$, $t \in \mathbb{R}_{\geq 0}$, generated by $-\mathbb{A}$ because it allows for an elegant way to represent solutions of parabolic equations
- (b) and the **fractional powers** \mathbb{A}^α , $\alpha \in \mathbb{R}$ because they allow to precisely link spatial smoothness to the powers of differential operators.

We will look at part (b) — i.e., fractional powers of operators — in section 3.5. Right now, we will deal with (a) by recalling a well-known result about sectorial operators and their correspondence to the generators of analytic semigroups:

Theorem 3.2.1. *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$ and let $M := \{\lambda \in \mathbb{C} : |\arg(\lambda)| < \psi\} \cup \{0\}$ for some $\psi \in (0, \frac{\pi}{2} - \phi)$. Then for all $z \in M \setminus \{0\}$ the following Bochner integral is well-defined:*

$$e^{-z\mathbb{A}} := \frac{1}{2\pi i} \int_{\Gamma} (\lambda - \mathbb{A})^{-1} e^{-z\lambda} d\lambda.$$

Here, Γ is an infinite curve in $S_{a,\phi} \subset \rho(\mathbb{A})$ which surrounds the spectrum $\sigma(\mathbb{A})$ of \mathbb{A} counter-clockwise in the sense that it lies entirely in some set $S_{a,\phi} \setminus S_{a',\phi}$, where $a' \in \mathbb{R}_{<a}$ - see figure 3.2 below for illustration. In that figure, an exemplary contour extending from $d + \infty e^{i\phi}$ to d to $d + \infty e^{-i\phi}$ for some $d < a$ is shown.

The integral does not depend on the specific choice of Γ and the following holds:

- (i) With the definition $e^{-0\mathbb{A}} := Id_H$, the operators $(e^{-z\mathbb{A}})_{z \in M}$ form an analytic semigroup on H with generator $-\mathbb{A}$.

(ii) The derivative of the mapping $M \setminus \{0\} \ni z \mapsto e^{-z\mathbb{A}} \in \mathcal{L}(H, H)$ is given by

$$\frac{d}{dz} [e^{-z\mathbb{A}}] = -\mathbb{A}e^{-z\mathbb{A}}.$$

(iii) For any analytic semigroup on H with generator $\mathbb{B} : D(\mathbb{B}) \subset H \rightarrow H$, the operator $-\mathbb{B}$ is sectorial.

See for example chapter 3 section 1 in [73] or Theorem 1.3.4. in [27] for proofs.

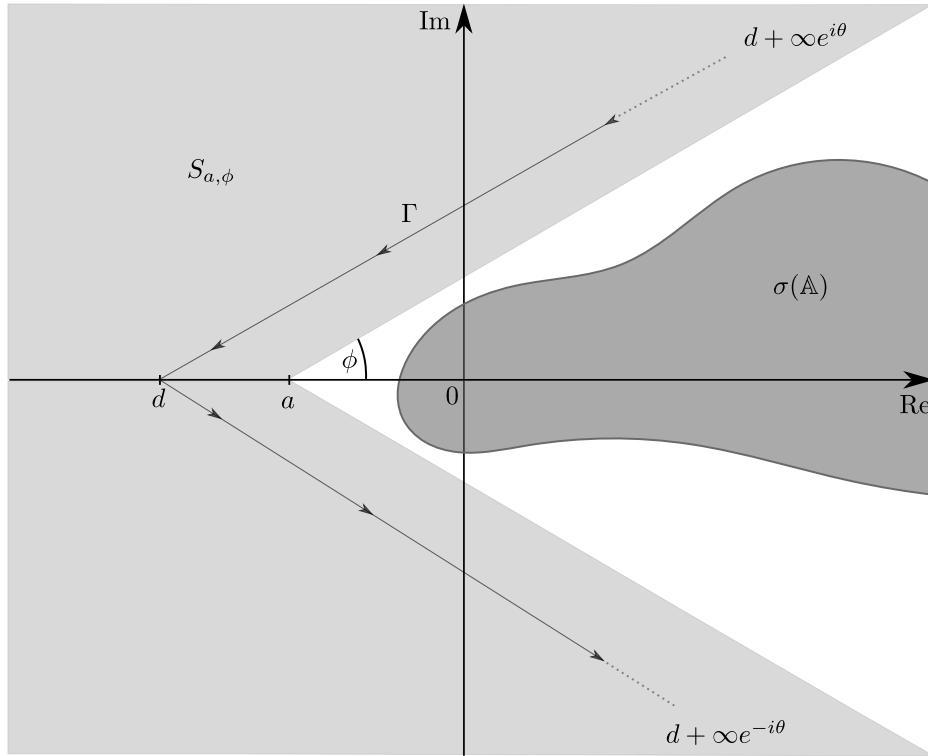


Figure 3.2: A contour Γ around the spectrum $\sigma(\mathbb{A})$ of a sectorial operator \mathbb{A} with sector $S_{a,\phi}$

To close this section, we introduce a tool that will be helpful when examining whether a given operator is sectorial or not.

Definition 3.2.4 (Numerical Range). *We define the numerical range $\Theta(\mathbb{B})$ of any linear operator $\mathbb{B} : D(\mathbb{B}) \subset H \rightarrow H$ as*

$$\Theta(\mathbb{B}) := \{(\mathbb{B}u, u)_H \in \mathbb{C} : u \in D(\mathbb{B}), \|u\|_H = 1\}.$$

Note that in the definition of the numerical range, we take elements of $D(\mathbb{B})$ but normalized in the norm of the larger space H .

The numerical range is useful in the theory of elliptic equations and also for showing sectoriality of operators — for example by applying the following proposition which is due to B.7. and B.8. in the appendix of [24].

Proposition 3.2.1. *Let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a linear, closed and densely-defined operator. Then it holds that:*

(i) *For all $\lambda \in \rho(\mathbb{A}) \setminus \overline{\Theta(\mathbb{A})}$ we have*

$$\|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(H,H)} \leq \frac{1}{\text{dist}\left(\lambda, \overline{\Theta(\mathbb{A})}\right)}. \quad (3.3)$$

(ii) *If $U \subset \mathbb{C} \setminus \overline{\Theta(\mathbb{A})}$ is open and connected and we also have $U \cap \rho(\mathbb{A}) \neq \emptyset$, then $U \subset \rho(\mathbb{A})$.*

Proof.

(i) Let $\lambda \in \rho(\mathbb{A}) \setminus \overline{\Theta(\mathbb{A})}$, so there is a $\delta > 0$ with $\delta = \text{dist}\left(\lambda, \overline{\Theta(\mathbb{A})}\right)$. For any $x \in D(\mathbb{A}) \setminus \{0\}$ we then get

$$\begin{aligned} \|x\|_H \|(\lambda - \mathbb{A})x\|_H &\geq |((\lambda - \mathbb{A})x, x)_H| \\ &= |\lambda\|x\|_H^2 - (\mathbb{A}x, x)_H| = \|x\|_H^2 \left| \lambda - \left(\mathbb{A} \frac{x}{\|x\|_H}, \frac{x}{\|x\|_H} \right)_H \right| \geq \|x\|_H^2 \delta \end{aligned}$$

so $\|x\|_H \leq \frac{1}{\delta} \|(\lambda - \mathbb{A})x\|_H$ which obviously holds for $x = 0$ as well. Since $\lambda \in \rho(\mathbb{A})$, $(\lambda - \mathbb{A})$ is invertible and we get $\|(\lambda - \mathbb{A})^{-1}y\|_H \leq \frac{1}{\delta} \|y\|_H$ for all $y \in H$.

(ii) This follows when looking a bit more closely at the resolvent mapping and writing it in a kind of Neumann series. A detailed proof can be found in B.7. and B.8. of the appendix in [24]. \square

This hints at a possible strategy for showing that a given closed operator is sectorial: narrow down the numerical range (which might be easier than finding the spectrum and/or the resolvent set) and then use the above proposition to get the resolvent bound.

3.3 Convection-Diffusion Operator

As a first application of the theory, we will look at this well-known standard example in the field of parabolic equations. Throughout this whole section we make the following assumptions: Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded domain with “sufficiently smooth” boundary (for example a convex polytope or a domain with boundary of class $C^{1,1}$ would suffice), let $\nu \in \mathbb{R}_{>0}$ be the (often very small) diffusion coefficient and let $b \in L^\infty(\Omega, \mathbb{R})^d$, $c \in L^\infty(\Omega, \mathbb{R})$.

Since we are interested in *spectral* properties of a convection-diffusion operator, we will work exclusively with complex-valued Sobolev spaces in this section. To shorten the presentation, we

omit the field \mathbb{C} from the notation, i.e., we use $L^2(\Omega) := L^2(\Omega, \mathbb{C})$, $H_0^1(\Omega) := H_0^1(\Omega, \mathbb{C})$ and so on.

We now introduce our convection-diffusion operator as follows:

Definition 3.3.1 (Convection-Diffusion Operator). *Let*

$$D(\mathbb{D}) := H^2(\Omega) \cap H_0^1(\Omega)$$

and let $\mathbb{D} : D(\mathbb{D}) \subset L^2(\Omega) \rightarrow L^2(\Omega)$ be the unbounded linear operator defined as

$$\mathbb{D}u := -\nu \Delta u + (b \cdot \nabla)u + cu \quad \text{for all } u \in D(\mathbb{D}).$$

Then we call \mathbb{D} the convection-diffusion operator with diffusion coefficient ν , transport direction b and zero-order coefficient c .

Note that the definition of the operator \mathbb{D} depends on parameters/coefficients that are omitted from the notation of the operator, i.e., we always assume that it is clear from the context, which parameters are used.

Of course it is possible to generalize this example tremendously by allowing varying coefficients also for the second derivatives or even allow solution dependent coefficients (which leads to quasilinear equations). We choose this basic setup because it simplifies the presentation but is still broad enough for many applications and retains most of its interesting features.

All derivatives here are meant as weak derivatives as usual. The choice of $D(\mathbb{D})$ ensures that they are indeed L^2 -functions — so \mathbb{D} is well-defined. In addition, we will later see, with the help of a well-known regularity result for elliptic equations, that this choice of $D(\mathbb{D})$ makes \mathbb{D} a closed operator.

We want to use this example to show in detail how to fit a well-known equation into this more abstract framework of sectorial operators and semigroups. To do this, we will recall some results from the theory of elliptic equations — mainly reproducing and using some results and proofs from [2]. In a first step we look at the associated sesquilinear form of the operator:

Definition 3.3.2 (Sesquilinear Form Associated with the Convection-Diffusion Operator). *Let*

$$\begin{aligned} B_{\mathbb{D}} : H_0^1(\Omega) \times H_0^1(\Omega) &\rightarrow \mathbb{C}, \\ (u, v) &\mapsto B_{\mathbb{D}}(u, v) := \nu (\nabla u, \nabla v)_{L^2} + ((b \cdot \nabla)u, v)_{L^2} + (cu, v)_{L^2}. \end{aligned}$$

Then we call $B_{\mathbb{D}}$ the sesquilinear form associated with the convection-diffusion operator.

Integration by parts quickly shows $(\mathbb{D}u, v)_{L^2} = B_{\mathbb{D}}(u, v)$ for all $u \in D(\mathbb{D})$ and $v \in H_0^1(\Omega)$. This means that for $u \in D(\mathbb{D})$ and $f \in L^2(\Omega)$ we have $\mathbb{D}u = f$ if and only if $B_{\mathbb{D}}(u, v) = (f, v)_{L^2}$ for all v in some dense subset of $H_0^1(\Omega)$. Later we will see that in many cases it is possible to neatly identify the domain of $B_{\mathbb{D}}$ — in this case $H_0^1(\Omega)$ — with the domain of $\mathbb{D}^{\frac{1}{2}}$, a fractional power of the operator that will be defined in section 3.5.

One of the most crucial properties of elliptic equations is given by the Garding inequality. That inequality is our main tool to show existence and uniqueness of solutions and sectoriality of the operator. It is basically a generalization of coercivity of the associated sesquilinear form.

Lemma 3.3.1 (Garding Inequality). *For all $u \in H_0^1(\Omega)$ we have*

$$\operatorname{Re} B_{\mathbb{D}}(u, u) \geq \frac{\nu}{2} \|\nabla u\|_{L^2}^2 - C_G \|u\|_{L^2}^2$$

where $C_G := \frac{1}{2\nu} \|b\|_{L^\infty}^2 + \|c\|_{L^\infty}$.

Proof. Let $u \in H_0^1(\Omega)$, then

$$\begin{aligned} \operatorname{Re} B_{\mathbb{D}}(u, u) &= \nu (\nabla u, \nabla u)_{L^2} + \operatorname{Re} ((b \cdot \nabla) u + cu, u)_{L^2} \\ &\geq \nu (\nabla u, \nabla u)_{L^2} - |((b \cdot \nabla) u + cu, u)_{L^2}| \\ &\geq \nu \|\nabla u\|_{L^2}^2 - \|b\|_{L^\infty} \|\nabla u\|_{L^2} \|u\|_{L^2} - \|c\|_{L^\infty} \|u\|_{L^2}^2 \\ &\geq \nu \|\nabla u\|_{L^2}^2 - \|b\|_{L^\infty} \left(\frac{\delta}{2} \|\nabla u\|_{L^2}^2 + \frac{1}{2\delta} \|u\|_{L^2}^2 \right) - \|c\|_{L^\infty} \|u\|_{L^2}^2. \end{aligned}$$

In the last step, we used — with some $\delta > 0$ — a version of Young's inequality. For $b \equiv 0$, the lemma obviously follows. Otherwise, we set $\delta := \frac{\nu}{\|b\|_{L^\infty}}$ which then proves the lemma as well. \square

Note that this looks very similar to a coercivity condition. For nontrivial b or c we do not get full coercivity. The strategy will be to shift the operator — and the associated sesquilinear form — to be able to use standard techniques like Lax-Milgram for the shifted equation.

Similarly to the numerical range $\Theta(\mathbb{D})$ of \mathbb{D} (see definition 3.2.4), we can also define a numerical range for its associated sesquilinear form: $\Theta(B_{\mathbb{D}}) := \{B_{\mathbb{D}}(u, u) \in \mathbb{C} : u \in H_0^1(\Omega), \|u\|_{L^2} = 1\}$. Using the conjugate symmetry of the dot product and the fact that $D(\mathbb{D}) \subset H_0^1(\Omega)$, we immediately see that $\bar{\lambda} \in \Theta(B_{\mathbb{D}})$ for any $\lambda \in \Theta(\mathbb{D})$. This numerical range $\Theta(B_{\mathbb{D}})$ enables us to formulate and prove the following existence and uniqueness result:

Theorem 3.3.1. *For all $\lambda \in \mathbb{C}$ that are **not** in $\overline{\Theta(B_{\mathbb{D}})}$ and all $f \in L^2(\Omega)$, there is a unique $u \in H_0^1(\Omega)$ so that*

$$\lambda (u, v)_{L^2} - B_{\mathbb{D}}(u, v) = (f, v)_{L^2}$$

for all $v \in H_0^1(\Omega)$. Additionally, that unique u fulfills the inequality

$$\|u\|_{H^1} \leq C_\lambda \|f\|_{L^2},$$

where $C_\lambda := \left(\frac{\frac{2}{\nu} (|\lambda| + C_G) + 1}{\operatorname{dist}(\lambda, \overline{\Theta(B_{\mathbb{D}})})} + \frac{2}{\nu} \right)$.

Proof. For a given $\lambda \in \mathbb{C} \setminus \overline{\Theta(B_{\mathbb{D}})}$, we set $B_\lambda(u, v) := \lambda (u, v)_{L^2} - B_{\mathbb{D}}(u, v)$ for all $u, v \in H_0^1(\Omega)$. For any $u \in H_0^1(\Omega) \setminus \{0\}$ we then get

$$\begin{aligned} |B_\lambda(u, u)| &= \left| \lambda \left(\frac{u}{\|u\|_{L^2}}, \frac{u}{\|u\|_{L^2}} \right)_{L^2} - B_{\mathbb{D}} \left(\frac{u}{\|u\|_{L^2}}, \frac{u}{\|u\|_{L^2}} \right) \right| \|u\|_{L^2}^2 \\ &\geq \operatorname{dist}(\lambda, \overline{\Theta(B_{\mathbb{D}})}) \|u\|_{L^2}^2 \end{aligned}$$

and thus with the Garding inequality

$$\begin{aligned}
 \|\nabla u\|_{L^2}^2 &\leq \frac{2}{\nu} (\operatorname{Re} B_{\mathbb{D}}(u, u) + C_G \|u\|_{L^2}^2) \\
 &\leq \frac{2}{\nu} (|B_{\lambda}(u, u)| + |\lambda(u, u)_{L^2}| + C_G \|u\|_{L^2}^2) \\
 &\leq \frac{2}{\nu} \left(|B_{\lambda}(u, u)| + \frac{|\lambda|}{\operatorname{dist}(\lambda, \overline{\Theta(B_{\mathbb{D}})})} |B_{\lambda}(u, u)| + \frac{C_G}{\operatorname{dist}(\lambda, \overline{\Theta(B_{\mathbb{D}})})} |B_{\lambda}(u, u)| \right).
 \end{aligned}$$

This then provides

$$\|u\|_{H^1}^2 \leq C_{\lambda} |B_{\lambda}(u, u)|.$$

Consequently, B_{λ} is a coercive sesquilinear form on $H_0^1(\Omega) \times H_0^1(\Omega)$. It is obviously continuous as well. A careful application of the Lax-Milgram theorem for sesquilinear forms — taking into account complex conjugates — then proves the theorem. For more details, see theorem 3.11 and corollary 3.12 in [2]. \square

As is well-known, if the boundary of Ω fulfills some smoothness conditions, we get even more regularity of solutions. We just quote the Result here, as the exact proof is rather technical and can be found in many textbooks on elliptic equations.

Theorem 3.3.2 (Elliptic Regularity). *Let $\partial\Omega$ be of class $C^{1,1}$ or let Ω be a convex polytope. Let $\lambda \in \mathbb{C}$, $f \in L^2(\Omega)$ and let there be a $u \in H_0^1(\Omega)$ so that $\lambda(u, v)_{L^2} - B_{\mathbb{D}}(u, v) = (f, v)_{L^2}$ for all $v \in H_0^1(\Omega)$.*

Then we have $u \in H^2(\Omega)$ with

$$\|u\|_{H^2} \leq C_{ER} (\|f\|_{L^2} + \|u\|_{L^2})$$

and the constant $C_{ER} > 0$ possibly depending on Ω , λ , c , b , ν but not on u or f .

For a proof with the assumption that $\partial\Omega$ is of class C^2 , see section 6.3 of [59]. For proofs in the cases that $\partial\Omega$ is of class $C^{1,1}$ or Ω is a convex polytope, see chapters 2. and 3. of [22].

If $u \in D(\mathbb{D})$, we obviously have $\mathbb{D}u \in L^2(\Omega)$ and — as explained above — with integration by parts we get $B_{\mathbb{D}}(u, v) = (\mathbb{D}u, v)_{L^2}$ for all $v \in H_0^1(\Omega)$. So the above theorem then gives us

$$\|u\|_{H^2} \leq C_{ER} (\|\mathbb{D}u\|_{L^2} + \|u\|_{L^2}). \quad (3.4)$$

Using this, we can now easily show that \mathbb{D} is closed.

Proposition 3.3.1. *The operator $\mathbb{D} : D(\mathbb{D}) \subset H \rightarrow H$ defined above is densely-defined and closed.*

Proof. Because $C_0^\infty(\Omega) \subset D(\mathbb{D})$ as well as $\overline{C_0^\infty(\Omega)}^{\|\cdot\|_{L^2}} = L^2(\Omega)$ — see theorem 3.1.2 — \mathbb{D} is densely-defined.

Now let $(u_k)_{k \in \mathbb{N}} \in (D(\mathbb{D}))^{\mathbb{N}}$ and $u \in D(\mathbb{D})$, $f \in L^2(\Omega)$ with

$$\lim_{k \rightarrow \infty} \|u_k - u\|_{L^2} = 0 \quad \text{and} \quad \lim_{k \rightarrow \infty} \|\mathbb{D}u_k - f\|_{L^2} = 0.$$

From (3.4) we then get

$$\begin{aligned} \|u_k - u_l\|_{H^2} &\leq C_{ER} (\|\mathbb{D}u_k - \mathbb{D}u_l\|_{L^2} + \|u_k - u_l\|_{L^2}) \\ &\leq C_{ER} (\|\mathbb{D}u_k - f\|_{L^2} + \|\mathbb{D}u_l - f\|_{L^2} + \|u_k - u_l\|_{L^2}) \xrightarrow{\min\{k,l\} \rightarrow \infty} 0. \end{aligned}$$

So $(u_k)_{k \in \mathbb{N}}$ is a Cauchy sequence in the Hilbert space $(D(\mathbb{D}), \|\cdot\|_{H^2})$ and thus we get

$$\lim_{k \rightarrow \infty} \|u_k - \tilde{u}\|_{H^2} = 0$$

for a $\tilde{u} \in D(\mathbb{D})$. Using $\|\cdot\|_{L^2} \leq \|\cdot\|_{H^2}$, we then get

$$\|u - \tilde{u}\|_{L^2} = \lim_{k \rightarrow \infty} \|u - u_k + u_k - \tilde{u}\|_{L^2} \leq \lim_{k \rightarrow \infty} (\|u - u_k\|_{L^2} + \|u_k - \tilde{u}\|_{H^2}) = 0,$$

so $u \in D(\mathbb{D})$. Lastly, using the facts that $\lim_{k \rightarrow \infty} \|u_k - u\|_{H^2} = 0$ and $\|\mathbb{D} \cdot\|_{L^2} \leq C' \|\cdot\|_{H^2}$ for some constant $C' > 0$ depending on the coefficients of \mathbb{D} , we get

$$\|f - \mathbb{D}u\|_{L^2} = \lim_{k \rightarrow \infty} \|f - \mathbb{D}u_k\|_{L^2} \leq C'' \lim_{k \rightarrow \infty} (\|f - \mathbb{D}u_k\|_{L^2} + \|u_k - u\|_{H^2}) = 0,$$

for another constant $C'' > 0$, so $\mathbb{D} : D(\mathbb{D}) \subset L^2(\Omega) \rightarrow L^2(\Omega)$ is closed.

Theorem 3.3.1 hints at a way to show sectoriality of \mathbb{D} with the help of the numerical range. In the following lemma, we examine the numerical range of \mathbb{D} and $B_{\mathbb{D}}$ a bit closer.

Lemma 3.3.2. *The numerical range $\Theta(B_{\mathbb{D}})$ of $B_{\mathbb{D}}$ is contained in the set*

$$U_{\mathbb{D}} := \left\{ \mu \in \mathbb{C} : \operatorname{Re} \mu \geq \frac{\nu}{2(C_P(\Omega))^2} - C_G, \operatorname{Im} \mu \leq \|b\|_{L^\infty} \sqrt{\frac{2}{\nu} (\operatorname{Re} \mu + C_G)} \right\},$$

where $C_G = \frac{1}{2\nu} \|b\|_{L^\infty} + \|c\|_{L^\infty}$ is the constant from the Garding inequality and $C_P(\Omega) > 0$ is the Poincaré constant.

Proof. Let $u \in H_0^1(\Omega)$ with $\|u\|_{L^2} = 1$. Then we have

$$\begin{aligned} |\operatorname{Im} B_{\mathbb{D}}(u, u)| &\leq |\operatorname{Im} (\nabla u, \nabla u)_{L^2}| + |\operatorname{Im} ((b \cdot \nabla) u, u)_{L^2}| + |\operatorname{Im} (cu, u)_{L^2}| \\ &= |\operatorname{Im} ((b \cdot \nabla) u, u)_{L^2}| \leq \|b\|_{L^\infty} \|\nabla u\|_{L^2} \|u\|_{L^2} \\ &= \|b\|_{L^\infty} \|\nabla u\|_{L^2} \leq \|b\|_{L^\infty} \sqrt{\frac{2}{\nu} (\operatorname{Re} B_{\mathbb{D}}(u, u) + C_G)}, \end{aligned}$$

and with the Poincaré inequality we get

$$\operatorname{Re} B_{\mathbb{D}}(u, u) + C_G \geq \frac{\nu}{2} \|\nabla u\|_{L^2}^2 \geq \frac{\nu}{2(C_P(\Omega))^2} \|u\|_{L^2}^2 = \frac{\nu}{2(C_P(\Omega))^2}.$$

□

Now we have everything we need to show that \mathbb{D} is a sectorial operator on $L^2(\Omega)$.

Theorem 3.3.3. *There exists $a \in \mathbb{R}$ and $\phi \in (0, \frac{\pi}{2})$ so that the above defined operator $\mathbb{D} : D(\mathbb{D}) \subset L^2(\Omega) \rightarrow L^2(\Omega)$ is a sectorial operator with sector $S_{a,\phi}$.*

Proof. The set $U_{\mathbb{D}}$ from the previous lemma is contained in a parabola with the vertex at $-C_G$. Thus, it is clear that we can pick $a \in \mathbb{R}_{\leq -C_G}$ and $\phi \in (0, \frac{\pi}{2})$ so that the sector $S_{a,\phi}$ is a subset of $\mathbb{C} \setminus \overline{U_{\mathbb{D}}}$. We wont go into the technical details of that here — see figure 3.3 below for illustration. The quantity $\text{dist}(S_{a,\phi}, \overline{U_{\mathbb{D}}})$ depends on the specific choice of a and ϕ . The larger that quantity is, the smaller the constant in the subsequent resolvent bounds will be.

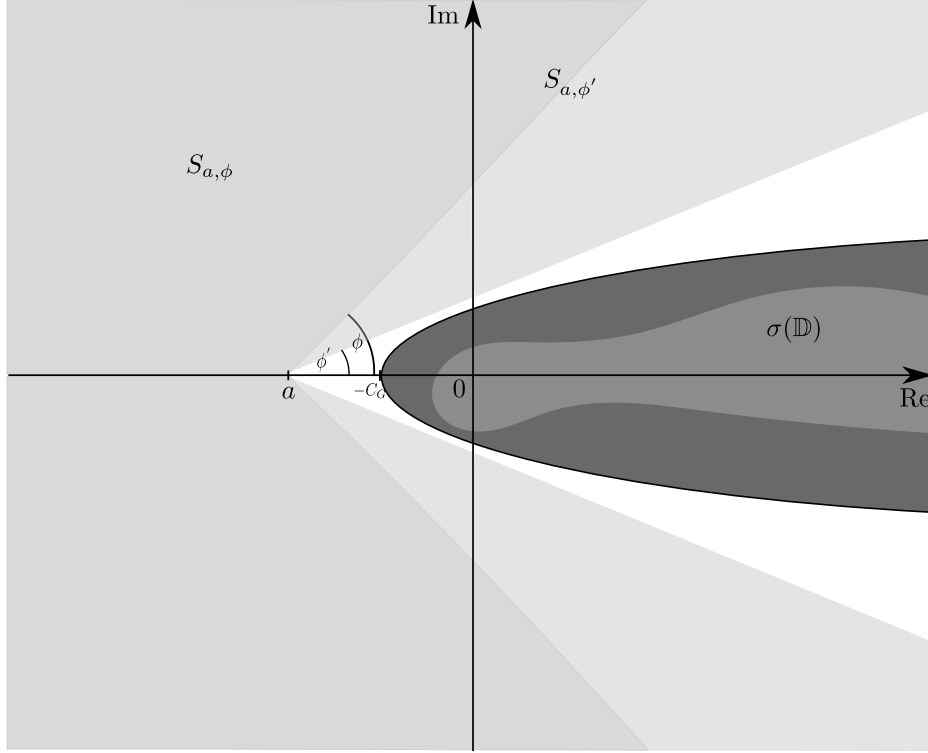


Figure 3.3: Two nested Sectors $S_{a,\phi} \subset S_{a,\phi'}$ and a parabolic area in $\mathbb{C} \setminus S_{a,\phi'}$ which contains the spectrum $\sigma(\mathbb{D})$ of a convection-diffusion operator \mathbb{D}

Now let $\lambda \in S_{a,\phi} \subset \mathbb{C} \setminus \overline{U_{\mathbb{D}}} \subset \mathbb{C} \setminus \overline{\Theta(B_{\mathbb{D}})}$. It then follows from theorems 3.3.1 and 3.3.2 that for each $f \in L^2(\Omega)$ there is a unique $u_f \in D(\mathbb{D})$ so that $(\lambda u_f - \mathbb{D}u_f, v) = \lambda(u_f, v)_{L^2} - B_{\mathbb{D}}(u_f, v) = (f, v)_{L^2}$ for all $v \in H_0^1(\Omega)$. Thus, the mapping $\lambda - \mathbb{D} : D(\mathbb{D}) \rightarrow L^2(\Omega)$, $u_f \mapsto f = (\lambda - \mathbb{D})u_f$ is a bijection. And using again theorem 3.3.1 we get for all $f \in L^2(\Omega)$:

$$\|(\lambda - \mathbb{D})^{-1}f\|_{L^2} \leq \|(\lambda - \mathbb{D})^{-1}f\|_{H^1} \leq C_{\lambda}\|f\|_{L^2},$$

so $(\lambda - \mathbb{D})^{-1} \in \mathcal{L}(L^2(\Omega), L^2(\Omega))$, i.e., $\lambda \in \rho(\mathbb{D})$.

Note that the constant

$$C_{\lambda} = \left(\frac{\frac{2}{\nu}(|\lambda| + C_G) + 1}{\text{dist}(\lambda, \overline{\Theta(B_{\mathbb{D}})})} + \frac{2}{\nu} \right)$$

in the bound above depends on λ in a way that does not directly give us the desired resolvent bound. Looking again at figure 3.3, it is clear that we can pick $0 < \phi' < \phi$, so that still $S_{a,\phi'} \subset \mathbb{C} \setminus \overline{U_{\mathbb{D}}}$. Some geometrical observations then reveal that

$$\text{dist}(\lambda, \overline{U_{\mathbb{D}}}) \geq \sin(\phi - \phi')|\lambda - a|.$$

Thus we can use proposition 3.2.1 to finally obtain for all $f \in L^2(\Omega)$ the bound

$$\|(\lambda - \mathbb{D})^{-1}\|_{\mathcal{L}(L^2, L^2)} \leq \frac{1}{\text{dist}(\lambda, \overline{\Theta(\mathbb{D})})} \leq \frac{1}{\text{dist}(\lambda, \overline{U_{\mathbb{D}}})} \leq \frac{1}{|\lambda - a| \sin(\phi - \phi')}.$$

□

3.4 Stokes Operator

Now we examine an operator which is similar to the one from the previous section. It is in some ways simpler because of its symmetry — and in some ways more complex because its description requires other function spaces than just the standard L^2 - or H^k -spaces.

The motivation to look at the Stokes operator is given, of course, by the incompressible Navier-Stokes equations, which can be rewritten in a form that involves said operator. See section 3.8 for a detailed look on how to formulate equations (2.13) and (2.14) into an initial value problem for an evolution equation that has no pressure term anymore and has the incompressibility condition ($\text{div } v = 0$) built into the function spaces. In the process of that transformation, the Stokes operator arises naturally.

Throughout this section let again $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded domain with “sufficiently smooth” boundary (for example a convex polytope or a domain with boundary of class $C^{1,1}$ would suffice). And even though we will see that the spectrum of the Stokes operator is contained within the positive real axis, we again work exclusively with *complex*-valued function spaces in this section.

We define several function spaces that are frequently used when working with the Navier-Stokes and Stokes equations. The following definitions and results are largely based on the well-known book by Temam [68].

Definition 3.4.1 (The Space $H_{\text{div}}(\Omega)$). *Let*

$$H_{\text{div}}(\Omega) := \{v \in L^2(\Omega)^d : \text{div } v \in L^2(\Omega)\}$$

and equip it with the dot product

$$(v, w)_{H_{\text{div}}} := (v, w)_{L^2} + (\text{div } v, \text{div } w)_{L^2}.$$

We can directly show the following:

Remark 3.4.1. $(H_{\text{div}}(\Omega), (\cdot, \cdot)_{H_{\text{div}}})$ is a Hilbert space.

Proof. One easily sees that the mapping $(\cdot, \cdot)_{H_{div}}$ really is a dot product. Now if $(v_k)_{k \in \mathbb{N}}$ is a Cauchy sequence in $H_{div}(\Omega)$, then $(v_k)_{k \in \mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega)^d$, thus converges to a $v \in L^2(\Omega)^d$ and $(\operatorname{div}(v_k))_{k \in \mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega)$, thus converges to a $w \in L^2(\Omega)$. Now let $\psi \in C_0^\infty(\Omega) \setminus \{0\}$ and $\epsilon > 0$. Then there is a $k \in \mathbb{N}$ so that we have

$$\|v_k - v\|_{L^2} + \|\operatorname{div}(v_k) - w\|_{L^2} < \frac{\epsilon}{\|\psi\|_{H^1}}.$$

Then we get

$$\begin{aligned} \left| \int_{\Omega} v \cdot \nabla \psi + w \psi \, d\Omega \right| &\leq \left| \int_{\Omega} (v - v_k) \cdot \nabla \psi \, d\Omega \right| + \left| \int_{\Omega} v_k \cdot \nabla \psi + w \psi \, d\Omega \right| \\ &\leq \|v_k - v\|_{L^2} \|\nabla \psi\|_{L^2} + \left| \int_{\Omega} (w - \operatorname{div} v_k) \psi \, d\Omega \right| \\ &\leq \|v_k - v\|_{L^2} \|\nabla \psi\|_{L^2} + \|\operatorname{div} v_k - w\|_{L^2} \|\psi\|_{L^2} < \epsilon. \end{aligned}$$

Since $\epsilon > 0$ was arbitrary, we have $\left| \int_{\Omega} v \cdot \nabla \psi + w \psi \, d\Omega \right| = 0$. Since this holds for all $\psi \in C_0^\infty(\Omega)$, we thus have $w = \operatorname{div} v \in L^2(\Omega)$, so $v \in H_{div}(\Omega)$. It is then very simple to show that $\|v_k - v\|_{H_{div}} \xrightarrow{k \rightarrow \infty} 0$. \square

The space $H_{div}(\Omega)$ is of importance to us because it allows for the definition of a trace operator that gives a useful characterization of the first of the two spaces which we will define next.

Definition 3.4.2 (The Spaces $H_\sigma(\Omega)$, $V_\sigma(\Omega)$ and $Q(\Omega)$). *Let*

$$H_\sigma(\Omega) := \overline{\{v \in C_0^\infty(\Omega)^d : \operatorname{div} v = 0\}}^{\|\cdot\|_{L^2}}$$

and equip it with the $L^2(\Omega)^d$ dot product. Let

$$V_\sigma(\Omega) := \overline{\{v \in C_0^\infty(\Omega)^d : \operatorname{div} v = 0\}}^{\|\cdot\|_{H^1}}$$

and equip it with the $H^1(\Omega)^d$ dot product. Finally, let

$$Q(\Omega) := \{p \in L^2(\Omega) : (p, 1)_{L^2} = 0\}$$

and equip it with the $L^2(\Omega)^d$ dot product.

Obviously, $H_\sigma(\Omega)$, $V_\sigma(\Omega)$ and $Q(\Omega)$ are Hilbert spaces. The latter of these spaces will be needed when the pressure is included in a formulation of the Navier-Stokes problem.

We want to define the Stokes operator as an unbounded operator on $H_\sigma(\Omega)$. To do that, we need more suitable characterizations of $H_\sigma(\Omega)$ and $V_\sigma(\Omega)$. As already mentioned above, we will use a trace operator on $H_{div}(\Omega)$ to characterize $H_\sigma(\Omega)$.

Theorem 3.4.1 (Trace Operator on $H_{div}(\Omega)$). *There exists a unique mapping $\gamma_n : H_{div}(\Omega) \rightarrow L^2(\partial\Omega)$ with the following properties:*

- (i) $\gamma_n \in L(H_{div}(\Omega), [H^{\frac{1}{2}}(\partial\Omega)]')$ where $H^{\frac{1}{2}}(\partial\Omega) \subset L^2(\Omega)$ is the image of the standard trace operator on $H^1(\Omega)$.

(ii) $\gamma_n(v) = v|_{\partial\Omega} \cdot n$ for all $v \in C_0^\infty(\bar{\Omega})^d$.

For a proof, see section 1.3. in [68]. We can now give the following results and characterizations for $H_\sigma(\Omega)$ and $V_\sigma(\Omega)$.

Theorem 3.4.2. *We have*

(i)

$$L^2(\Omega)^d = H_\sigma(\Omega) \oplus \{u \in L^2(\Omega)^d : u = \nabla q, q \in H^1(\Omega)\}$$

(ii) as well as

$$V_\sigma(\Omega) = \{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0\}$$

(iii) and

$$H_\sigma(\Omega) = \{v \in L^2(\Omega)^d : \gamma_n(v) = 0, \operatorname{div} v = 0\}.$$

Proof. We will give just an outline here. For details, see section 1.4. in [68]. To show (i), we examine for a given $u \in L^2(\Omega)^d$ the problems

$$\begin{aligned} p &\in H_0^1(\Omega) \text{ so that } \Delta p = \operatorname{div} u \text{ on } \Omega \\ q &\in H^1(\Omega) \text{ so that } \Delta q = 0 \text{ on } \Omega \text{ and } \nabla q \cdot n = u - \nabla p \text{ on } \partial\Omega \end{aligned}$$

are by the theory of elliptic equations uniquely solvable (q is unique up to a constant). Setting $v := u - \nabla p + \nabla q$ directly gives $\operatorname{div} v = 0$.

We want to remark here, that the solutions p and q of the above problems have for $k \in \{0, 1, 2, \dots\}$ and $u \in H^k(\Omega)^d$ by the theory of elliptic equations the regularity

$$\|p\|_{H^{k+1}} \leq C_1 \|\operatorname{div} u\|_{H^{k-1}} \leq C_1 \|u\|_{H^k}$$

and

$$\|q\|_{H^{k+1}} \leq C_2 \|u - \nabla p\|_{H^k} \leq (C_2 + C_1 C_2) \|u\|_{H^k}$$

with constants $C_1, C_2 > 0$ that do not depend on u . For $u \in L^2(\Omega)$ we have $\operatorname{div} u \in H^{-1}(\Omega)$ of course. Thus we can deduce

$$\|v\|_{H^k} = \|u - \nabla p + \nabla q\|_{H^k} \leq C_3 \|u\|_{H^k}$$

with a constant $C_3 > 0$ independent of u . This will be used to show boundedness of a projection to $H_\sigma(\Omega)$ that we will define below the proof of this theorem.

Finally, through integration by parts we obtain that the spaces $H_\sigma(\Omega)$ and

$\{u \in L^2(\Omega)^d : u = \nabla q, q \in H^1(\Omega)\}$ are orthogonal.

We will now show (ii) and (iii). The inclusions

$$V_\sigma(\Omega) \subset \{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0\}$$

and

$$H_\sigma(\Omega) \subset \{v \in L^2(\Omega)^d : \gamma_n(v) = 0, \operatorname{div} v = 0\}$$

are shown through the density of $C_0^\infty(\Omega)$ in $V_\sigma(\Omega)$ and $H_\sigma(\Omega)$, respectively, and with similar arguments as were used in the proof of remark 3.4.1. In the proof of the latter of the two inclusions, the continuity of the trace operator γ_n is used to show that $\gamma_n(v) = 0$ is indeed true for v in $H_\sigma(\Omega)$.

Now, a deep and important result based on a theorem by de Rham states that for any $l \in H^{-1}(\Omega)^d$, which fulfills $\langle l, v \rangle = 0$ for all $v \in C_0^\infty(\Omega)^d \cap V_\sigma(\Omega)$, there is a $p \in Q$ with $\langle l, v \rangle = -(p, \operatorname{div} l)_{L^2}$ for all $v \in H_0^1(\Omega)^d$. Further down in theorem 3.4.3, we will present an often used variant of this fact. For more details on the formulation we just stated, see section 1.4. in [68]. Using that formulation, we immediately obtain that any $l \in H^{-1}(\Omega)^d$ which vanishes on $V_\sigma(\Omega)$ also vanishes on $\{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0\}$. A standard result from Banach space theory then provides that $V_\sigma(\Omega)$ is dense in $\{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0\}$ and thus $V_\sigma(\Omega) = \{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0\}$.

Lastly, let X be the orthogonal complement of $H_\sigma(\Omega)$ in $\{v \in L^2(\Omega)^d : \gamma_n(v) = 0, \operatorname{div} v = 0\}$. Then by (i) we have

$$X \subset \{v \in L^2(\Omega)^d : \gamma_n(v) = 0, \operatorname{div} v = 0\} \cap \{v \in L^2(\Omega)^d : v = \nabla q, q \in H^1(\Omega)\}.$$

This quickly reveals $X = \{0\}$ in the following way:

$v \in X$ means $\gamma_n(v) = 0$, $\operatorname{div} v = 0$ and $v = \nabla p$ with $p \in H^1(\Omega)$. This means p solves the Neumann problem $\Delta p = \operatorname{div} v = 0$, $\nabla p \cdot n = v \cdot n = \gamma_n(v) = 0$. Thus p is a constant and $v = \nabla p = 0$. \square

By (i) of the last theorem, we get the following L^2 -projection to $H_\sigma(\Omega)$:

Definition 3.4.3 (Leray Projection). *Define the mapping $P_\sigma : L^2(\Omega)^d \rightarrow H_\sigma(\Omega)$ by setting for any $u \in L^2(\Omega)^d$*

$$P_\sigma(u) := v,$$

where $v \in H_\sigma(\Omega)$, $q \in H^1(\Omega)$ with $u = v + \nabla q$ (see theorem 3.4.2 (i)). We call P_σ the Leray projection.

We quickly notice the following:

Remark 3.4.2. *Let $k \in \{0, 1, 2, \dots\}$. From the proof of part (i) of theorem 3.4.2 we directly deduce that P_σ continuously maps $H^k(\Omega)^d$ into $H^k(\Omega)^d \cap H_\sigma(\Omega)$. I.e. there is a constant $C_k > 0$ so that for all $u \in H^k(\Omega)^d$ we have*

$$\|P_\sigma u\|_{H^k} \leq C_k \|u\|_{H^k}.$$

Before we finally define the Stokes operator and show that it fits into our framework of sectorial operators, we want to give a precise definition for the gradient as an operator and examine some of its properties. The gradient operator is particularly useful for understanding different formulations (with or without pressure) of Stokes, Navier Stokes and Oseen-type problems.

Definition 3.4.4 (Gradient Operator). *Let*

$$V_\sigma^\perp(\Omega) := \{l \in H^{-1}(\Omega)^d : l(v) = 0 \text{ for all } v \in V_\sigma(\Omega)\}$$

and for all $p \in Q(\Omega) = \{p \in L^2(\Omega) : (p, 1)_{L^2} = 0\}$ define

$$\langle \text{grad } p, v \rangle := -(p, \text{div } v)_{L^2} \quad \text{for all } v \in H_0^1(\Omega)^d.$$

Next we state an important result for the gradient operator, which involves the well-known inf-sup condition.

Theorem 3.4.3 (Inf-Sup Condition). *For all $p \in Q(\Omega)$ we have $\text{grad } p \in V_\sigma^\perp(\Omega)$ and the following statements are true:*

(i) *As a mapping with domain $Q(\Omega)$ and codomain $V_\sigma^\perp(\Omega)$, the above defined gradient operator*

$$\text{grad} : Q(\Omega) \rightarrow V_\sigma^\perp(\Omega)$$

is an isomorphism.

(ii) *There exists a constant $\gamma > 0$, so that for all $p \in Q(\Omega)$*

$$\|\text{grad } p\|_{H^{-1}} \geq \gamma \|p\|_{L^2}.$$

(iii) *The inf-sup condition holds, i.e., there exists a $\gamma > 0$ so that*

$$\inf_{p \in Q(\Omega) \setminus \{0\}} \sup_{v \in H_0^1(\Omega)^d \setminus \{0\}} \frac{|(p, \text{div } v)_{L^2}|}{\|p\|_{L^2} \|\nabla v\|_{L^2}} \geq \gamma.$$

A proof can be found in section 2.2.3 of [54]. A common way to partly prove this result is to show that the three statements are equivalent, and then use the fact that one of them is true — such as statement (ii), which is sometimes called *Nečas inequality*.

As a first application of this, we take a look at two different variational formulations of a stationary Stokes problem. We show how to recover the unique pressure corresponding to the *pressure-less* solution to the variational formulation in divergence-free spaces.

Remark 3.4.3. *Let $f \in H^{-1}(\Omega)^d$ and $v \in V_\sigma(\Omega)$ so that*

$$(\nabla v, \nabla w)_{L^2} = \langle f, w \rangle \quad \text{for all } w \in V_\sigma(\Omega). \quad (3.5)$$

Then the mapping

$$l : H_0^1(\Omega)^d \rightarrow \mathbb{C}, w \mapsto (\nabla v, \nabla w)_{L^2} - \langle f, w \rangle$$

fulfills $l \in V_\sigma^\perp(\Omega)$ so that we can use part (i) of the above theorem to obtain a unique pressure $p \in Q(\Omega)$ with $l = \text{grad } p$. Thus, these $v \in H_0^1(\Omega)^d$ and $p \in Q$ fulfill the following variational formulation as well:

$$(\nabla v, \nabla w)_{L^2} - (p, \text{div } w)_{L^2} = \langle f, w \rangle \quad \text{for all } w \in H_0^1(\Omega)^d, \quad (3.6)$$

$$(\text{div } v, q)_{L^2} = 0 \quad \text{for all } q \in Q(\Omega). \quad (3.7)$$

It is not hard to see that in fact (3.5) is equivalent to (3.6), (3.7).

With the help of the Leray projection introduced in definition 3.4.3 we can now define the Stokes operator. Later in section 3.8 we will also use that projection to state the entire incompressible Navier-Stokes problem emerging from (2.13), (2.14) as a problem on the spaces $H_\sigma(\Omega)$ and $V_\sigma(\Omega)$.

Definition 3.4.5 (Stokes Operator). *Let*

$$D(\mathbb{S}) := V_\sigma(\Omega) \cap H^2(\Omega)^d$$

and let $\mathbb{S} : D(\mathbb{S}) \subset H_\sigma(\Omega) \rightarrow H_\sigma(\Omega)$ be the unbounded linear operator defined as

$$\mathbb{S}v := -P_\sigma(\Delta v) \quad \text{for all } v \in D(\mathbb{S}).$$

Then we call \mathbb{S} the Stokes operator.

Obviously this operator is well-defined and we will promptly show that it fits into our framework of sectorial operators. Similarly to the convection-diffusion operator, we will use the Lax Milgram theorem in the proof. And again — just like for the convection-diffusion operator — we will make use of higher regularity than the Lax-Milgram theorem provides us with.

Theorem 3.4.4 (Stokes Regularity). *Let $\nu \in \mathbb{R}_{>0}$ and let $\partial\Omega$ be of class $C^{1,1}$ or let Ω be a convex polytope. Let $f \in L^2(\Omega)$ and let there be a $v \in H_0^1(\Omega)^d$ and a $p \in Q(\Omega) = \{q \in L^2(\Omega) : (q, 1)_{L^2} = 0\}$ so that*

$$\begin{aligned} \nu(\nabla v, \nabla w)_{L^2} - (p, \text{div } w)_{L^2} &= (f, w)_{L^2} \\ (\text{div } v, q)_{L^2} &= 0 \end{aligned}$$

for all $w \in H_0^1(\Omega)^d$ and all $q \in Q(\Omega)$. Then we have $v \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$ with

$$\|v\|_{H^2} + \|p\|_{H^1} \leq C_{SR}\|f\|_{L^2}$$

with a constant $C_{SR} > 0$ that depends on ν and Ω but not on v , p or f .

For a proof with the assumption that $\partial\Omega$ is of class C^2 , see proposition 2.2 in [68]. For proofs in the cases that $\partial\Omega$ is of class $C^{1,1}$ or Ω is a convex polytope, see theorem 6.3 in [14].

We now have everything we need, to show that \mathbb{S} is sectorial.

Theorem 3.4.5. *There exist $a \in \mathbb{R}_{>0}$ and $w \in (0, \frac{\pi}{2})$ so that the above defined operator \mathbb{S} is a sectorial operator on $H_\sigma(\Omega)$ with sector $S_{a,w}$. \mathbb{S} is also self-adjoint.*

Proof. By Definition, we have $V_\sigma(\Omega) = \overline{\{v \in C_0^\infty(\Omega)^d : \operatorname{div} v = 0\}}^{\|\cdot\|_{H^1}}$. Since obviously $\{v \in C_0^\infty(\Omega)^d : \operatorname{div} v = 0\} \subset H^2(\Omega)^d$, we immediately get $\overline{D(\mathbb{S})}^{\|\cdot\|_{L^2}} = H_\sigma(\Omega)$ by definition as well. Thus, \mathbb{S} is densely defined.

Next, we show that \mathbb{S} is symmetric. Let $v, w \in D(\mathbb{S})$. Then in particular $v, w \in H_0^1(\Omega)$ and $v, w \in H_\sigma(\Omega)$, so they vanish on the boundary and

$$\begin{aligned} (P_\sigma z, w)_{L^2} &= (z, w)_{L^2}, \\ (v, P_\sigma z)_{L^2} &= (v, z)_{L^2} \end{aligned}$$

for all $z \in L^2(\Omega)^d$. Thus we get

$$\begin{aligned} (-P_\sigma(\Delta v), w)_{L^2} &= (-\Delta v, w)_{L^2} = (\nabla v, \nabla w)_{L^2} \\ &= (v, -\Delta w)_{L^2} = (v, -P_\sigma(\Delta w))_{L^2}. \end{aligned} \quad (3.8)$$

Now we show that \mathbb{S} is a bijection. Just as we did for the convection-diffusion operator, we will define a sesquilinear form associated with \mathbb{S} :

$$B_\mathbb{S} : V_\sigma(\Omega) \times V_\sigma(\Omega) \rightarrow \mathbb{C}, \quad (v, w) \mapsto B_\mathbb{S}(v, w) := (\nabla v, \nabla w)_{L^2}.$$

Similar to (3.8) above we get $(\mathbb{S}v, w)_{L^2} = B_\mathbb{S}(v, w)$ for all $v \in D(\mathbb{S})$ and $w \in V_\sigma(\Omega)$. This means that for $v \in D(\mathbb{S})$ and $f \in H_\sigma(\Omega)$ we have $\mathbb{S}v = f$ if and only if $B_\mathbb{S}(v, w) = (f, w)_{L^2}$ for all w in some dense subset of $V_\sigma(\Omega)$. Analogous to the convection-diffusion problem, we will see that it is possible to neatly identify the domain of $B_\mathbb{S} - V_\sigma(\Omega)$ — with the domain of $\mathbb{S}^{\frac{1}{2}}$, a fractional power of the operator that will be defined in section 3.5.

$B_\mathbb{S}$ is obviously coercive. For any $f \in H_\sigma(\Omega) \subset [V_\sigma(\Omega)]'$ the Lax-Milgram theorem then provides us with a unique $v \in V_\sigma(\Omega)$ so that $B_\mathbb{S}(v, w) = (f, w)_{L^2}$ for all $w \in V_\sigma(\Omega)$.

From remark 3.4.3 we know that there exists a unique pressure $p \in \{q \in L^2(\Omega) : (q, 1)_{L^2} = 0\}$ with

$$(\nabla v, \nabla w)_{L^2} - (p, \operatorname{div} w)_{L^2} = (f, w)_{L^2} \quad \text{for all } w \in H_0^1(\Omega)^d$$

and thus we conclude with theorem 3.4.4 that

$$\|v\|_{H^2} \leq C_{SR} \|f\|_{L^2} \quad (3.9)$$

for some constant C_{SR} independent of v and f .

So indeed, for any $f \in L^2(\Omega)^d$ there exists a unique $v \in D(\mathbb{S})$ with $B_\mathbb{S}(v, w) = (f, w)_{L^2}$ for all $w \in V_\sigma(\Omega)$, i.e., $\mathbb{S}v = f$. Thus, \mathbb{S} is a bijection.

By (3.9) (or alternatively the Lax-Milgram theorem) the inverse \mathbb{S}^{-1} is in $\mathcal{L}(H_\sigma(\Omega), H_\sigma(\Omega))$:

$$\|\mathbb{S}^{-1}f\|_{L^2} \leq \|\mathbb{S}^{-1}f\|_{H^2} \leq C_{SR} \|f\|_{L^2} \quad \text{for all } f \in H_\sigma(\Omega).$$

Thus we have $0 \in \rho(\mathbb{S})$.

Next, we want to show that \mathbb{S} is self adjoint. First of all we notice that \mathbb{S}^{-1} is symmetric as

well:

Let $f, g \in H_\sigma(\Omega)$, then there exist $v, w \in D(\mathbb{S})$ with $\mathbb{S}^{-1}f = v$ and $\mathbb{S}^{-1}g = w$ and we get

$$(\mathbb{S}^{-1}f, g)_{L^2} = (v, \mathbb{S}w)_{L^2} = (\mathbb{S}v, w)_{L^2} = (f, \mathbb{S}^{-1}g)_{L^2}.$$

An easy to show consequence of remark 3.1.2 and the symmetry of \mathbb{S}^{-1} is that $D(\mathbb{S}^{-1}) \subset D((\mathbb{S}^{-1})^*) \subset H_\sigma(\Omega)$ and $\mathbb{S}^{-1}f = (\mathbb{S}^{-1})^*f$ for all $f \in D(\mathbb{S}^{-1})$. So because $D(\mathbb{S}^{-1}) = H_\sigma(\Omega)$, we get $D(\mathbb{S}^{-1}) = D((\mathbb{S}^{-1})^*)$ and thus $\mathbb{S}^{-1} = (\mathbb{S}^{-1})^*$. From B.4. b) in the appendix of [24] — or from proposition 3.5.3 further below, which is based on results in [24] — we get $(\mathbb{S}^{-1})^* = (\mathbb{S}^*)^{-1}$ and thus

$$\mathbb{S} = (\mathbb{S}^{-1})^{-1} = ((\mathbb{S}^{-1})^*)^{-1} = ((\mathbb{S}^*)^{-1})^{-1} = \mathbb{S}^*.$$

Since self-adjoint operators are in particular closed — see again remark 3.1.2 — all we have left to do now, is to prove that the spectrum of \mathbb{S} is contained in a sector within the open right half plane of \mathbb{C} and that \mathbb{S} admits to a sectoriality bound on that sector. To do this, we will first look at the numerical range $\Theta(\mathbb{S})$ of \mathbb{S} . Let $v \in D(\mathbb{S}) \subset H_0^1(\Omega)^d$ with $\|v\|_{H_\sigma} = 1$. Using (3.8) and the Poincaré inequality, we get:

$$(v, \mathbb{S}v)_{L^2} = (\nabla v, \nabla v)_{L^2} = \|\nabla v\|_{L^2}^2 \geq C_P(\Omega)\|v\|_{L^2}^2 = C_P(\Omega) > 0.$$

With the Poincaré constant $C_P(\Omega)$. It follows that $\overline{\Theta(\mathbb{S})} \subset \mathbb{R}_{\geq C_P(\Omega)} \subset \mathbb{R}_{>0}$. Now choose any $a \in (0, C_P(\Omega))$ and $\phi \in (0, \frac{\pi}{2})$. The specific choice of ϕ will influence the constants in the resolvent bounds. As we showed above, we have $0 \in \rho(\mathbb{S})$. Obviously, $0 \in S_{a,\phi}$ and $S_{a,\phi} \subset \mathbb{C} \setminus \overline{\Theta(\mathbb{S})}$ hold as well. Part (ii) of proposition 3.2.1 then provides us with $S_{a,\phi} \subset \rho(\mathbb{S})$ and part (i) of 3.2.1 gives us

$$\|(\lambda - \mathbb{S})^{-1}\|_{L^2} \leq \frac{1}{\text{dist}(\lambda, \overline{\Theta(\mathbb{S})})}$$

for all $\lambda \in S_{a,\phi}$. It is a bit technical, but not particularly difficult to show for all $\lambda \in S_{a,\phi}$ that $\text{dist}(\lambda, \mathbb{R}_{\geq C_P(\Omega)}) \geq |\lambda - a| \sin(\phi)$. Combining everything, we get for all $\lambda \in S_{a,\phi}$:

$$\|(\lambda - \mathbb{S})^{-1}\|_{L^2} \leq \frac{1}{\text{dist}(\lambda, \overline{\Theta(\mathbb{S})})} \leq \frac{1}{\text{dist}(\lambda, \mathbb{R}_{\geq C_P(\Omega)})} \leq \frac{1}{|\lambda - a| \sin(\phi)}.$$

□

3.5 Fractional Spaces

After these examples, we will return again to the abstract framework on a separable and complex Hilbert space H . Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$.

We start this section by looking a bit more closely at the situation that the complete negative real axis and a neighborhood of 0 are contained in the resolvent set of A . It is easy to see that this is the case if and only if $a > 0$. As we saw above, this is true for the Stokes operator but not necessarily for the convection-diffusion operator (depending on b and c).

Before we define general (fractional) powers of operators, we observe that for all $k \in \mathbb{N}$, the operator

$$\mathbb{A}^k = \underbrace{\mathbb{A} \cdot \mathbb{A} \cdot \dots \cdot \mathbb{A}}_{k\text{-times}}$$

is well-defined with $D(\mathbb{A}^k) = \{x \in D(\mathbb{A}) : \mathbb{A}x \in D(\mathbb{A}^{k-1})\}$ for all $k \in \mathbb{N}$ with $k \geq 2$. If \mathbb{A} is a sectorial operator with sector $S_{a,\phi}$ and $a > 0$, then \mathbb{A}^{-1} exists, and so $\mathbb{A}^{-k} = (\mathbb{A}^k)^{-1}$ is a well-defined linear, bounded and injective operator on H with $\mathcal{R}(\mathbb{A}^{-k}) = D(\mathbb{A}^k)$ for all $k \in \mathbb{N}$.

Now, we will define **fractional** powers of sectorial operators with a strictly positive spectrum. Similar to the exponential function of sectorial operators (analytic semigroup) we will do this with the help of an integral along a contour around the spectrum.

Theorem 3.5.1 (Negative Fractional Powers). *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$ and $a > 0$. Then for all $\alpha > 0$ the following Bochner integral is well-defined:*

$$\mathbb{A}^{-\alpha} := \frac{1}{2\pi i} \int_{\Gamma} (\lambda - \mathbb{A})^{-1} \lambda^{-\alpha} d\lambda.$$

Here, Γ is a contour in $S_{a,\phi} \cap \{z \in \mathbb{C} : \operatorname{Re} z > 0\}$ which surrounds the spectrum $\sigma(\mathbb{A})$ of \mathbb{A} counterclockwise in the sense that it lies entirely in some set $S_{a,\phi} \setminus S_{a',\phi}$, where $a' \in (0, a)$. The situation is similar to the one depicted in figure 3.2, with the difference that in our case here, the spectrum $\sigma(\mathbb{A})$ of \mathbb{A} and the contour Γ lie entirely within the right half plane. Because of $a > 0$ this kind of contour exists.

The integral does not depend on the specific choice of Γ and the following assertions hold:

- (i) For all $\alpha > 0$ we have $\mathbb{A}^{-\alpha} \in \mathcal{L}(H, H)$ and $\mathcal{N}(\mathbb{A}^{-\alpha}) = \{0\}$.
- (ii) For all $k \in \mathbb{N}$, \mathbb{A}^{-k} defined in this way agrees with the previous definition for natural numbers as expected.
- (iii) For all $\alpha, \epsilon > 0$ we have $(\epsilon \mathbb{A})^{-\alpha} = \epsilon^{-\alpha} \mathbb{A}^{-\alpha}$.
- (iv) For all $\alpha, \beta > 0$ we have $\mathbb{A}^{-\alpha} \mathbb{A}^{-\beta} = \mathbb{A}^{-(\alpha+\beta)} = \mathbb{A}^{-\beta} \mathbb{A}^{-\alpha}$.
- (v) The function $\mathbb{R}_{>0} \ni \alpha \mapsto \mathbb{A}^{-\alpha} \in \mathcal{L}(H, H)$ is continuous and $\lim_{\alpha \rightarrow 0^+} \|\mathbb{A}^{-\alpha} - \operatorname{Id}_H\|_{\mathcal{L}(H, H)} = 0$.

For a proof see for example chapter 2 section 7 in [73]. Based on the above theorem, we can now define positive powers as well.

Definition 3.5.1 (Positive Fractional Powers). *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$ and $a > 0$. Then we define*

$$\begin{aligned} \mathbb{A}^0 &:= \operatorname{Id}_H, \\ \mathbb{A}^\alpha &:= (\mathbb{A}^{-\alpha})^{-1}, \quad D(\mathbb{A}^\alpha) := \mathcal{R}(\mathbb{A}^{-\alpha}), \quad \text{for all } \alpha \in \mathbb{R}_{>0}. \end{aligned}$$

As we can see, one has to be a bit more careful with positive powers because of the varying domains. Negative powers do not have that problem, as they always lead to fully defined, bounded operators. We can, however, say the following:

Proposition 3.5.1. *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$ and $a > 0$. Then the following holds:*

- (i) *For all $\alpha, \beta \in \mathbb{R}$ with $\beta \leq \alpha$ we have $D(\mathbb{A}^\alpha) \subseteq D(\mathbb{A}^\beta)$.*
- (ii) *For all $\alpha \in \mathbb{R}$, \mathbb{A}^α is closed and densely defined.*
- (iii) *For all $\alpha, \beta \in \mathbb{R}$ we have $\mathbb{A}^\alpha \mathbb{A}^\beta x = \mathbb{A}^\beta \mathbb{A}^\alpha x = \mathbb{A}^{\alpha+\beta} x$ for all $x \in D(\mathbb{A}^\gamma)$, where $\gamma := \max\{\alpha, \beta, \alpha + \beta\}$.*

Proof.

- (i) For $0 < \beta < \alpha$ we have $\mathbb{A}^{-\alpha} = \mathbb{A}^{-\beta} \mathbb{A}^{-(\alpha-\beta)}$ by theorem 3.5.1, so $\mathcal{R}(\mathbb{A}^{-\alpha}) \subseteq \mathcal{R}(\mathbb{A}^{-\beta})$ and consequently $D(\mathbb{A}^\alpha) \subseteq D(\mathbb{A}^\beta)$.
- (ii) Let $\alpha \in \mathbb{R}_{>0}$, $(x_n)_{n \in \mathbb{N}} \in (D(\mathbb{A}^\alpha))^\mathbb{N}$ and $x, y \in H$ with $\|x_n - x\|_H \xrightarrow{n \rightarrow \infty} 0$ and $\|\mathbb{A}^\alpha x_n - y\|_H \xrightarrow{n \rightarrow \infty} 0$. Then, since $\mathbb{A}^{-\alpha}$ is bounded, we have $\|x_n - \mathbb{A}^{-\alpha} y\|_H \xrightarrow{n \rightarrow \infty} 0$ and consequently $\mathbb{A}^{-\alpha} y = x$ so $x \in D(\mathbb{A}^\alpha)$ and $\mathbb{A}^\alpha x = y$, thus \mathbb{A}^α is closed.

To show that $D(\mathbb{A}^\alpha)$ is dense in H , a bit more work is required. At this point we repeat the proof from pages 84 and 95 in [73] to illustrate some arguments when dealing with sectorial operators and their powers. For all $n \in \mathbb{N}$, define $J_n : H \rightarrow H$ via $J_n := (Id + n^{-1}\mathbb{A})^{-1}$. First, we observe that

$$\|J_n\|_{\mathcal{L}(H,H)} = \|-n(-n - \mathbb{A})^{-1}\|_{\mathcal{L}(H,H)} \leq M_{\mathbb{A}} \frac{n}{|n+a|} \leq M_{\mathbb{A}}$$

with the sectoriality constant $M_{\mathbb{A}} > 0$, so J_n is bounded independent of n . We then have for all $n, k \in \mathbb{N}$:

$$R(J_n^k) = D\left((Id + n^{-1}\mathbb{A})^k\right) \subseteq D(\mathbb{A}^k).$$

Now let $k \in \mathbb{N}$ with $k \geq \alpha$. If we assume for a moment that J_n^k converges strongly to Id_H , i.e., for all $u \in H$, we have $\|u - J_n^k u\|_H \xrightarrow{n \rightarrow \infty} 0$, the density of $D(\mathbb{A}^\alpha)$ in H immediately follows:

For a given $\epsilon > 0$ and $u \in H$, we can pick $n \in \mathbb{N}$ large enough so that for

$$u_\epsilon := J_n^k u \in R(J_n^k) \subset D(\mathbb{A}^k) \stackrel{a)}{\subset} D(\mathbb{A}^\alpha)$$

we get $\|u - u_\epsilon\|_H < \epsilon$.

It remains to show that J_n^k converges strongly to Id_H . First of all, we observe that

$$J_n = Id - \mathbb{A}(n + \mathbb{A})^{-1},$$

so

$$\begin{aligned} \|u - J_n u\|_H &= \|\mathbb{A}(n + \mathbb{A})^{-1} u\|_H = \|(n + \mathbb{A})^{-1} \mathbb{A} u\|_H \\ &\leq \|-(-n - \mathbb{A})^{-1}\|_{\mathcal{L}(H,H)} \|\mathbb{A} u\|_H \leq \frac{M_{\mathbb{A}}}{|n+a|} \|\mathbb{A} u\|_H \xrightarrow{n \rightarrow \infty} 0 \end{aligned}$$

for all $u \in D(\mathbb{A})$. Now let $u \in H$. Since $D(\mathbb{A})$ is dense in H , we can choose for a given $\epsilon > 0$ a $u_\epsilon \in D(\mathbb{A})$ with $\|u - u_\epsilon\|_H < \epsilon$ and then a $n_0 \in \mathbb{N}$ so that $\|u_\epsilon - J_n u_\epsilon\|_H < \epsilon$ for all $n \in \mathbb{N}$ with $n \geq n_0$. For those n we have

$$\|u - J_n u\|_H \leq \|u - u_\epsilon\|_H + \|u_\epsilon - J_n u_\epsilon\|_H + \|J_n u_\epsilon - J_n u\|_H \leq (2 + M_{\mathbb{A}})\epsilon,$$

so J_n converges to Id on the whole of H . Finally we get for $k \in \mathbb{N}$:

$$\begin{aligned} \|u - J_n^k u\|_H &\leq \|u - J_n u\|_H + \|J_n u - J_n^k u\|_H \\ &\leq \|u - J_n u\|_H + M_{\mathbb{A}} \|u - J_n^{k-1} u\|_H \leq \dots \leq \|u - J_n u\|_H \sum_{l=0}^{k-1} M_{\mathbb{A}}^l \xrightarrow{n \rightarrow \infty} 0. \end{aligned}$$

Thus proving that J_n^k converges strongly to Id_H .

(iii) This follows from part (iv) of theorem 3.5.1. \square

Before we come to the main focus of this section, let us notice that for a given sectorial operator \mathbb{A} on $S_{a,\phi}$ with a not necessarily positive, we can always look at the operator $\mathbb{A} - a + \epsilon$ for any $\epsilon > 0$. That operator then obviously has the same dense domain as \mathbb{A} . In addition, it is easily seen that $\mathbb{A} - a + \epsilon$ is closed and even sectorial with sector $S_{\epsilon,\phi}$, so fractional powers can be defined. As we will see, the specific choice of epsilon is largely irrelevant. To show that, we need the following important result about the graph norms of fractional powers of operators.

Theorem 3.5.2. *Let $\mathbb{A}_1 : D(\mathbb{A}_1) \subset H \rightarrow H$ be a sectorial operator on H with sector S_{a_1,ϕ_1} , $a_1 > 0$, and let $\mathbb{A}_2 : D(\mathbb{A}_2) \subset H \rightarrow H$ be a sectorial operator on H with sector S_{a_2,ϕ_2} , $a_2 > 0$. Suppose that $D(\mathbb{A}_1) = D(\mathbb{A}_2)$ with $\frac{1}{C_{1,2}} \|\mathbb{A}_1 u\|_H \leq \|\mathbb{A}_2 u\|_H \leq C_{1,2} \|\mathbb{A}_1 u\|_H$ for some constant $C_{1,2} > 0$ and all $u \in D(\mathbb{A}_1)$. Further assume that there exists a $\beta \in [0, 1)$ and a constant $C_\beta > 0$ so that*

$$\|(\mathbb{A}_1 - \mathbb{A}_2)u\|_H \leq C_\beta \|\mathbb{A}_1^\beta u\|_H \quad \text{for all } u \in D(\mathbb{A}_1). \quad (3.10)$$

Then for each $\alpha \in [0, 1]$ we have $D(\mathbb{A}_1^\alpha) = D(\mathbb{A}_2^\alpha)$ and there exists a constant $C > 0$, so that

$$\frac{1}{C} \|\mathbb{A}_1^\alpha u\|_H \leq \|\mathbb{A}_2^\alpha u\|_H \leq C \|\mathbb{A}_1^\alpha u\|_H$$

for all $u \in D(\mathbb{A}_1^\alpha) = D(\mathbb{A}_2^\alpha)$, i.e., the graph norms of \mathbb{A}_1^α and \mathbb{A}_2^α are equivalent. The constant C depends on the sectoriality constants $M_{\mathbb{A}_1}, M_{\mathbb{A}_2}$ of \mathbb{A}_1 and \mathbb{A}_2 , on a_1, ϕ_1, a_2, ϕ_2 , on $C_{1,2}$, on C_β and on β

A proof of this can be constructed from theorem 2.26 in [73] in conjunction with an idea from the proof of theorem 1.4.6 in [27] which shows that even though condition (3.10) is not symmetric, that condition is sufficient to obtain the symmetric result when exchanging the roles of \mathbb{A}_1 and \mathbb{A}_2 .

This now lets us define fractional spaces.

Definition 3.5.2 (Fractional Spaces). *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$ and let $\epsilon > 0$. Then we define for any $\alpha \in [0, 1]$ the fractional space*

$$H_\alpha(\mathbb{A}) := D((\mathbb{A} - a + \epsilon)^\alpha)$$

and equip it with the dot product

$$H_\alpha(\mathbb{A}) \times H_\alpha(\mathbb{A}) \ni (u, v) \mapsto (u, v)_{\mathbb{A}^\alpha} := ((\mathbb{A} - a + \epsilon)^\alpha u, (\mathbb{A} - a + \epsilon)^\alpha v)_H.$$

The induced norm is denoted by $\|\cdot\|_{\mathbb{A}^\alpha}$.

Applying the above theorem 3.5.2 with $\beta = 0$ to $\mathbb{A}_1 := \mathbb{A} - a + \epsilon_1$ and $\mathbb{A}_2 := \mathbb{A} - a + \epsilon_2$ for any $\epsilon_1, \epsilon_2 > 0$ shows that different choices of ϵ all lead to the same space and dot products that induce equivalent norms. Thus, it is reasonable to suppress the dependence on the specific choice of ϵ in the definition above. This way, we are able to define meaningful fractional spaces even for sectorial operators that do not only have eigenvalues with strictly positive real part — like the convection-diffusion operator.

We note the following results about fractional spaces:

Proposition 3.5.2. *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$.*

- (i) *For all $\alpha \in [0, 1]$, $H_\alpha(\mathbb{A})$ is a Hilbert space.*
- (ii) *For all $1 \geq \alpha \geq \beta \geq 0$, $H_\alpha(\mathbb{A})$ is a dense subspace of $H_\beta(\mathbb{A})$ with continuous embedding.*

Proof. Let $1 \geq \alpha > \beta > 0$, $\epsilon > 0$ and $\tilde{\mathbb{A}} := \mathbb{A} - a + \epsilon$.

- (i) It is elementary to show that the mapping $H_\alpha(\mathbb{A}) \times H_\alpha(\mathbb{A}) \ni (u, v) \mapsto (u, v)_{\mathbb{A}^\alpha}$ really is a dot product. Now let $(u_n)_{n \in \mathbb{N}}$ be a Cauchy sequence in $H_\alpha(\tilde{\mathbb{A}}) \subset H$, i.e., $(\tilde{\mathbb{A}}^\alpha u_n)_{n \in \mathbb{N}}$ is a Cauchy sequence in H . Because H is complete, there is a $w \in H$ with $\|\tilde{\mathbb{A}}^\alpha u_n - w\|_H \xrightarrow{n \rightarrow \infty} 0$. Since

$$\|u_n - u_m\|_H = \|\tilde{\mathbb{A}}^{-\alpha} \tilde{\mathbb{A}}^\alpha (u_n - u_m)\|_H \leq \|\tilde{\mathbb{A}}^{-\alpha}\|_H \|\tilde{\mathbb{A}}^\alpha (u_n - u_m)\|_H \xrightarrow{n, m \rightarrow \infty} 0,$$

$(u_n)_{n \in \mathbb{N}}$ is also a Cauchy sequence in H , so there is a $u \in H$ with $\|u_n - u\|_H \xrightarrow{n \rightarrow \infty} 0$. Because $\tilde{\mathbb{A}}^\alpha$ is closed, we get $u \in D(\tilde{\mathbb{A}}^\alpha)$ and $\tilde{\mathbb{A}}^\alpha u = w$ so

$$\|u_n - u\|_{\tilde{\mathbb{A}}^\alpha} = \|\tilde{\mathbb{A}} u_n - \tilde{\mathbb{A}} u\|_H \xrightarrow{n \rightarrow \infty} 0.$$

- (ii) Let $u \in D(\tilde{\mathbb{A}}^\beta)$ so $\tilde{\mathbb{A}}^\beta u \in H$. Since $D(\tilde{\mathbb{A}}^{\alpha-\beta})$ is dense in H , there is a sequence $(u_n)_{n \in \mathbb{N}} \in (D(\tilde{\mathbb{A}}^{\alpha-\beta}))^\mathbb{N}$ with $\|u_n - \tilde{\mathbb{A}}^{\alpha-\beta} u\|_H \xrightarrow{n \rightarrow \infty} 0$. For all $n \in \mathbb{N}$ we have $u_n \in D(\tilde{\mathbb{A}}^{\alpha-\beta})$ so obviously $\tilde{\mathbb{A}}^{-\beta} u_n \in D(\tilde{\mathbb{A}}^\alpha)$. Overall we get

$$\|\tilde{\mathbb{A}}^{-\beta} u_n - u\|_{\tilde{\mathbb{A}}^\alpha} = \|u_n - \tilde{\mathbb{A}}^\beta u\|_H \xrightarrow{n \rightarrow \infty} 0.$$

Similarly, we see that

$$\|u\|_{\tilde{\mathbb{A}}^\beta} = \|\tilde{\mathbb{A}}^{\beta-\alpha} \tilde{\mathbb{A}}^\alpha u\|_H \leq \|\tilde{\mathbb{A}}^{-(\alpha-\beta)}\|_{\mathcal{L}(H, H)} \|u\|_{\tilde{\mathbb{A}}^\alpha}$$

for all $u \in D(\tilde{\mathbb{A}}^\alpha)$, so the embedding is continuous. \square

Even though we do not plan use the following result, we want to mention it for the sake of completeness. For a proof see for example theorem 2.23 in [73].

Theorem 3.5.3. *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$, $a > 0$, and let $0 < \alpha < 1$. Then \mathbb{A}^α is a sectorial operator on H with sector $S_{a',\phi'}$ with $\phi' \leq \alpha\phi$ and some suitable $a' > 0$.*

To conclude this section, we gather some related facts about the adjoint of sectorial operators. These results are taken from [24] — more specifically from proposition 4.5. as well as B.4 and B.5 of the appendix of that work.

Proposition 3.5.3. *Let \mathbb{A} be a sectorial operator on H with sector $S_{a,\phi}$ and $a > 0$. Then*

(i) $\mathbb{A}^* : D(\mathbb{A}^*) \rightarrow H$ is a sectorial operator on H with sector $S_{a,\phi}$.

(ii) $(\mathbb{A}^\alpha)^* = (\mathbb{A}^*)^\alpha$ for all $\alpha \in \mathbb{R}$.

Overall, we thus see that fractional powers of operators do behave largely as would be expected. The laws of exponents hold and exponentiation can be exchanged with the adjoint. We do, of course, have to be careful with the domains of fractional powers. Though from proposition 3.5.2 we obtain that those domains are neatly nested Hilbert spaces with continuous embeddings.

3.5.1 Examples of Fractional Spaces

Here, we will apply the concept of fractional powers to the convection-diffusion and Stokes operators. To start, we define the Dirichlet Laplacian and the vector-valued Dirichlet Laplacian. Just as in sections 3.3 and 3.4, let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded Domain with “sufficiently smooth” boundary and let all function spaces be *complex*-valued.

Definition 3.5.3 (Laplacian). *Let*

$$\begin{aligned} D(\mathbb{L}) &:= H_0^1(\Omega) \cap H^2(\Omega) \\ D(\vec{\mathbb{L}}) &:= H_0^1(\Omega)^d \cap H^2(\Omega)^d \end{aligned}$$

and let $\mathbb{L} : D(\mathbb{L}) \subset L^2(\Omega) \rightarrow L^2(\Omega)$ and $\vec{\mathbb{L}} : D(\vec{\mathbb{L}}) \subset L^2(\Omega)^d \rightarrow L^2(\Omega)^d$ be the unbounded linear operators defined as

$$\begin{aligned} \mathbb{L}u &:= -\Delta u \quad \text{for all } u \in D(\mathbb{L}), \\ \vec{\mathbb{L}}u &:= -\Delta u \quad \text{for all } u \in D(\vec{\mathbb{L}}), \end{aligned}$$

with the symbol Δ , of course, being applied in the scalar sense for \mathbb{L} and in the vector sense for $\vec{\mathbb{L}}$. Then we call \mathbb{L} the Dirichlet Laplacian and $\vec{\mathbb{L}}$ the vector-valued Dirichlet Laplacian.

Notice how the Dirichlet Laplacian is equal to the convection-diffusion operator \mathbb{D} for the case that there is no convection ($b \equiv 0$), no reaction ($c \equiv 0$) and the diffusion coefficient is equal to 1. Thus, we immediately get that the Dirichlet Laplacian is a sectorial operator.

One easily sees — very similar to the proof for the Stokes operator — that \mathbb{L} is self-adjoint and that its spectrum is contained in $\mathbb{R}_{\geq c}$ for some $c > 0$. This can also be shown for the vector-valued Dirichlet Laplacian with basically the same arguments.

Now we want to show a well-known identification of some fractional spaces associated with these operators as Sobolev spaces. We only look at the fraction $\frac{1}{2}$ here, since that is all we need in this work. Furthermore, that case can be handled with elementary methods and without delving into the theory of interpolation spaces. We want to stress, though, that far more general identification results for fractional spaces are available — see for example [19], [21] or [23] for more information.

Proposition 3.5.4. *Let \mathbb{S} be the Stokes operator defined as in section 3.4 and let $\mathbb{L}, \vec{\mathbb{L}}$ be the Dirichlet Laplacian and vector-valued Dirichlet Laplacian defined above. Then the following equalities hold with norm equivalence:*

(i)

$$H_{\frac{1}{2}}(\mathbb{L}) = H_0^1(\Omega).$$

(ii)

$$H_{\frac{1}{2}}(\vec{\mathbb{L}}) = H_0^1(\Omega)^d.$$

(iii)

$$H_{\frac{1}{2}}(\mathbb{S}) = V_\sigma(\Omega).$$

Proof. We only prove (iii) here — the other assertions can be shown similarly.

First of all we note that by definition 3.4.2 $D(\mathbb{S}) = H^2(\Omega)^d \cap V_\sigma(\Omega)$ is a dense subset of $V_\sigma(\Omega)$. From proposition 3.5.2 we get that $H_{\frac{1}{2}}(\mathbb{S})$ is a Hilbert space of which $D(\mathbb{S})$ is also a dense subset.

Using the L^2 -orthogonality of the Leray projection and the self-adjointness of the Stokes operator alongside proposition 3.5.3, we get for all $v \in D(\mathbb{S})$ that

$$\|\nabla v\|_{L^2}^2 = (\mathbb{S}v, v)_{L^2} = (\mathbb{S}^{\frac{1}{2}}v, \mathbb{S}^{\frac{1}{2}}v)_{L^2} = \|v\|_{\mathbb{S}^{\frac{1}{2}}}^2. \quad (3.11)$$

With the Poincaré inequality we thus see that on $D(\mathbb{S})$ the norms $\|\cdot\|_{H^1}$ and $\|\cdot\|_{\mathbb{S}^{\frac{1}{2}}}$ are equivalent. Now let $u \in V_\sigma(\Omega)$. Then there is a sequence $(u_n)_{n \in \mathbb{N}} \in (D(\mathbb{S}))^\mathbb{N}$ so that

$$\lim_{n \rightarrow \infty} \|u - u_n\|_{H^1} = 0.$$

Obviously, $(u_n)_{n \in \mathbb{N}}$ is a Cauchy sequence in $V_\sigma(\Omega)$. Because of (3.11) it is then also a Cauchy sequence in $H_{\frac{1}{2}}(\mathbb{S})$, so there is a $w \in H_{\frac{1}{2}}(\mathbb{S})$ with

$$\lim_{n \rightarrow \infty} \|w - u_n\|_{\mathbb{S}^{\frac{1}{2}}} = 0.$$

This means that for any $\epsilon > 0$ there is an $N \in \mathbb{N}$ large enough so that for all $n \in \mathbb{N}$ with $n \geq N$ we have both $\|u - u_n\|_{H^1} < \epsilon$ and $\|w - u_n\|_{\mathbb{S}^{\frac{1}{2}}} < \epsilon$. Using again proposition 3.5.2, we see that the L^2 -norm can be bounded by the $H_{\frac{1}{2}}(\mathbb{S})$ -norm. Hence we get the following for all $\epsilon > 0$ and all $n \in \mathbb{N}$ that are small enough:

$$\|u - w\|_{L^2} \leq \|u - u_n\|_{H^1} + \|u_n - w\|_{\mathbb{S}^{\frac{1}{2}}} \leq \epsilon.$$

Thus, we have $u = w$ and consequently $V_\sigma(\Omega) \subset H_{\frac{1}{2}}(\mathbb{S})$. Furthermore, for all $\epsilon > 0$ and all $n \in \mathbb{N}$ that are small enough we also get:

$$\|u\|_{\mathbb{S}^{\frac{1}{2}}} \leq \epsilon + \|u_n\|_{\mathbb{S}^{\frac{1}{2}}} \leq \epsilon + C\|u_n - u + u\|_{H^1} \leq \epsilon + C\epsilon + C\|u\|_{H^1}$$

with some constant independent of u and ϵ . This means that the $H_{\frac{1}{2}}(\mathbb{S})$ -norm is bounded by the H^1 -norm.

It is easy to see that the continuous inclusion $H_{\frac{1}{2}}(\mathbb{S}) \subset V_\sigma(\Omega)$ can be shown with the same arguments.

Now we want to show that the asymmetry in the convection-diffusion operator does not have significant influence on the fractional spaces produced by that operator.

Remark 3.5.1. *Let \mathbb{D} be the convection-diffusion operator defined as in section 3.3 with the diffusion coefficient $\nu > 0$, the convection direction $b \in L^\infty(\Omega, \mathbb{R})^d$ and the reaction coefficient $c \in L^\infty(\Omega, \mathbb{R})$. Then we have with norm equivalence:*

$$H_s(\mathbb{D}) = H_s(\mathbb{L}), \quad \text{for all } s \in [0, 1],$$

In particular we have $H_{\frac{1}{2}}(\mathbb{D}) = H_0^1(\Omega)$.

Proof. Let \mathbb{L} be the Dirichlet Laplacian defined above and set $\tilde{\mathbb{L}} := \nu\mathbb{L}$ as well as $\tilde{\mathbb{D}} := \mathbb{D} + d$ for some $d \in \mathbb{R}$ which is chosen so that $\tilde{\mathbb{D}}$ is sectorial on some sector that contains 0. From the previous results in section 3.5 we quickly get $H_\alpha(\tilde{\mathbb{L}}) = H_\alpha(\mathbb{L})$ and $H_\alpha(\tilde{\mathbb{D}}) = H_\alpha(\mathbb{D})$ with equivalent norms for all $\alpha \in [0, 1]$. By definition we have $D(\tilde{\mathbb{L}}) = D(\tilde{\mathbb{D}}) = H^2(\Omega) \cap H_0^1(\Omega)$.

Now we apply theorem 3.5.2 with $\alpha = \frac{1}{2}$, $\mathbb{A}_1 = \tilde{\mathbb{L}}$ and $\mathbb{A}_2 = \tilde{\mathbb{D}}$. If all requirements of that theorem are fulfilled, it provides $H_\beta(\tilde{\mathbb{L}}) = H_\beta(\tilde{\mathbb{D}})$ with norm equivalence for all $\beta \in [0, 1]$, thus proving our claim as follows:

$$H_\beta(\mathbb{L}) = H_\beta(\tilde{\mathbb{L}}) = H_\beta(\tilde{\mathbb{D}}) = H_\beta(\mathbb{D})$$

with norm equivalence for all $\beta \in [0, 1]$.

Now we show that all requirements of theorem 3.5.2 are fulfilled.

The Lax-Milgram theorem provides us for all $u \in H^2(\Omega) \cap H_0^1(\Omega)$ with the estimates

$$\|u\|_{H^1} \leq C_{\tilde{\mathbb{L}}} \|\tilde{\mathbb{L}}u\|_{L^2} \quad \text{and} \quad \|u\|_{H^1} \leq C_{\tilde{\mathbb{D}}} \|\tilde{\mathbb{D}}u\|_{L^2},$$

where the constant $C_{\tilde{\mathbb{L}}} > 0$ depends on ν and the Poincaré constant $C_P(\Omega)$ and $C_{\tilde{\mathbb{D}}} > 0$ also depends on both of those but additionally on $\|b\|_{L^\infty}$ and $\|c\|_{L^\infty}(\Omega)$.

This proves that for all $u \in H^2(\Omega) \cap H_0^1(\Omega)$ we have

$$\begin{aligned} \|\tilde{\mathbb{L}}u\|_{L^2} &\leq \|-\nu\Delta u + (b \cdot \nabla)u + (c + d)u\|_{L^2} + \|(b \cdot \nabla)u + (c + d)u\|_{L^2} \\ &\leq \|\tilde{\mathbb{D}}u\|_{L^2} + (\|b\|_{L^\infty} + \|c\|_{L^\infty} + |d|)\|u\|_{H^1} \\ &\leq (1 + C_{\tilde{\mathbb{D}}}(\|b\|_{L^\infty} + \|c\|_{L^\infty} + |d|))\|\tilde{\mathbb{D}}u\|_{L^2} \end{aligned}$$

and similarly

$$\|\tilde{\mathbb{D}}u\|_{L^2} \leq (1 + C_{\tilde{\mathbb{L}}}(\|b\|_{L^\infty} + \|c\|_{L^\infty} + |d|))\|\tilde{\mathbb{L}}u\|_{L^2}.$$

Using part (i) of proposition 3.5.4, we show the following for all $u \in H^2(\Omega) \cap H_0^1(\Omega)$:

$$\begin{aligned} \|(\tilde{\mathbb{L}} - \tilde{\mathbb{D}})u\|_{L^2} &\leq \|(b \cdot \nabla)u + (c + d)u\|_{L^2} \\ &\leq (\|b\|_{L^\infty} + \|c\|_{L^\infty} + |d|)\|u\|_{H^1} \\ &\leq \nu^{-\frac{1}{2}}C_{\tilde{\mathbb{L}}, \frac{1}{2}}(\|b\|_{L^\infty} + \|c\|_{L^\infty} + |d|)\|\tilde{\mathbb{L}}^{\frac{1}{2}}u\|_{L^2}, \end{aligned}$$

where $C_{\tilde{\mathbb{L}}, \frac{1}{2}} > 0$ is the constant originating from part (i) of proposition 3.5.4. Thus, the proof is completed. \square

3.6 Extended Operators

We now show how to extend a sectorial operator, that has some additional properties, to a bounded operator on a triplet of spaces in a natural way. The ideas for this construction are based on the beginning of section 2. in [42]. As usual we assume that H denotes a separable complex Hilbert space.

Proposition 3.6.1. *Let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a sectorial operator on H with sector $S_{a, \phi}$ and let $\tilde{\mathbb{A}} := \mathbb{A} - a + \epsilon$ with some arbitrary $\epsilon > 0$. In addition, let $V := H_{\frac{1}{2}}(\mathbb{A}) = H_{\frac{1}{2}}(\mathbb{A}^*)$ and*

$$C_{\mathbb{A}}\|\tilde{\mathbb{A}}^{\frac{1}{2}}u\|_H \leq \|\tilde{\mathbb{A}}^{*\frac{1}{2}}u\|_H \leq \frac{1}{C_{\mathbb{A}}}\|\tilde{\mathbb{A}}^{\frac{1}{2}}u\|_H$$

for all $u \in V$, i.e., the associated norms of the spaces are equivalent. When V is equipped with the dot product

$$V \times V \ni (u, v) \mapsto (u, v)_V := \left(\tilde{\mathbb{A}}^{\frac{1}{2}}u, \tilde{\mathbb{A}}^{\frac{1}{2}}v \right)_H,$$

the following holds:

(i) (V, H, V') forms a triplet of spaces.

(ii) \mathbb{A} has a well-defined extension $\mathbb{A}_V : V \rightarrow V'$ in $\mathcal{L}(V, V')$ that is given for each $u \in V$ by

$$\langle \mathbb{A}_V u, v \rangle := (\tilde{\mathbb{A}}^{\frac{1}{2}}u, \tilde{\mathbb{A}}^{*\frac{1}{2}}v)_H + (a - \epsilon)(u, v)_H$$

for all $v \in V$.

If $a > 0$ in the above proposition, we can just set $\epsilon := a$ and get $\tilde{\mathbb{A}} = \mathbb{A}$ as well as $\langle \mathbb{A}_V u, v \rangle = (\mathbb{A}^{\frac{1}{2}}u, \mathbb{A}^{*\frac{1}{2}}v)_H$ for all $u, v \in V$.

Proof.

(i) From proposition 3.5.2 it follows that $(V, (\cdot, \cdot)_V)$ is a Hilbert space and from part (ii) of that proposition, we get the dense and continuous inclusion $V \xhookrightarrow{d} H$.

- (ii) We first check $\mathbb{A}_V \in \mathcal{L}(V, V')$. \mathbb{A}_V is obviously linear. By (i), there exists a constant $C_{GT} > 0$ so that $\|v\|_H \leq C_{GT}\|v\|_V$ for all $v \in V$. Using that and the equivalence of the norms $\|\tilde{\mathbb{A}}^{\frac{1}{2}} \cdot\|_H$ and $\|\tilde{\mathbb{A}}^{*\frac{1}{2}} \cdot\|_H$ on V , we get for all $u, v \in V$

$$\begin{aligned} |\langle \mathbb{A}_V u, v \rangle| &= \left| (\tilde{\mathbb{A}}^{\frac{1}{2}} u, \tilde{\mathbb{A}}^{*\frac{1}{2}} v)_H + (a - \epsilon)(u, v)_H \right| \\ &\leq \|\tilde{\mathbb{A}}^{\frac{1}{2}} u\|_H \|\tilde{\mathbb{A}}^{*\frac{1}{2}} v\|_H + |a - \epsilon| \|u\|_H \|v\|_H \leq \left(\frac{1}{C_{\mathbb{A}}} + C_{GT} |a - \epsilon| \right) \|u\|_V \|v\|_V \end{aligned}$$

so $\mathbb{A}_V \in \mathcal{L}(V, V')$.

Now we show, in what way \mathbb{A}_V is an extension of \mathbb{A} . Using the embeddings given in (i), we obtain for each $u \in D(\mathbb{A})$ that $\mathbb{A}u \in V'$ with $\langle \mathbb{A}u, v \rangle = (\mathbb{A}u, v)_H$ for each $v \in V$. On the other hand, we have for all $u \in D(\mathbb{A})$ and $v \in V$

$$\langle \mathbb{A}_V u, v \rangle = \left(\tilde{\mathbb{A}}^{\frac{1}{2}} u, \tilde{\mathbb{A}}^{*\frac{1}{2}} v \right)_H + (a - \epsilon)(u, v)_H = (\tilde{\mathbb{A}}u, v)_H + (a - \epsilon)(u, v)_H = (\mathbb{A}u, v)_H,$$

so $\mathbb{A}_V|_{D(\mathbb{A})} = \mathbb{A}$.

This also shows that the extension does not depend on the specific choice of ϵ and is thus well-defined. To see that, pick another $\hat{\epsilon} > 0$ and denote the corresponding extension by $\hat{\mathbb{A}}_V$. Then we have $\hat{\mathbb{A}}_V|_{D(\mathbb{A})} = \mathbb{A} = \mathbb{A}_V|_{D(\mathbb{A})}$ so the operators $\hat{\mathbb{A}}_V, \mathbb{A}_V \in \mathcal{L}(V, V')$ coincide on a dense subspace of V . Thus, they are equal. \square

From now on we will always write \mathbb{A} instead of \mathbb{A}_V when using the above defined extension. If $a > 0$, we can preserve invertability of the extension and show some strengthened resolvent bounds as we will see in the following theorem.

Theorem 3.6.1. *Let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a sectorial operator on H with sector $S_{a, \phi}$, $a > 0$ and sectoriality constant $M_{\mathbb{A}} > 0$ and let $V := H_{\frac{1}{2}}(\mathbb{A}) = H_{\frac{1}{2}}(\mathbb{A}^*)$ with equivalent norms as in the proposition above. Furthermore, for all $\lambda \in \mathbb{C}$ we define the operator $\lambda : V \rightarrow V'$, in accordance with the usual embeddings within triplets of spaces, as*

$$\langle \lambda v, w \rangle := (\lambda v, w)_H$$

for all $v, w \in V$. Obviously, that operator is continuous.

For all $\delta \in \mathbb{R}_{<a}$ and all $\lambda \in S_{\delta, \phi}$ we then get:

- (i) $(\lambda - \mathbb{A}) : D(\mathbb{A}) \rightarrow H$ is a bijection with

$$\begin{aligned} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(H, H)} &\leq \frac{M_1}{1 + |\lambda|}, \\ \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(V, V)} &\leq \frac{M_1}{1 + |\lambda|}, \end{aligned}$$

where $M_1 > 0$ can be chosen to depend only on δ , a , ϕ and $M_{\mathbb{A}}$.

- (ii) $(\lambda - \mathbb{A}) : V \rightarrow V'$ is a continuous bijection with

$$\begin{aligned} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(V', V')} &\leq \frac{M_2}{1 + |\lambda|}, \\ \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(V', V)} &\leq M_3, \end{aligned}$$

where $M_2, M_3 > 0$ can be chosen to depend only on $\delta, a, \phi, M_{\mathbb{A}}$ and upper bounds for $\|\mathbb{A}\|_{\mathcal{L}(V, V')}, \|\mathbb{A}^{-1}\|_{\mathcal{L}(V', V)}$.

Proof. Let $\delta \in \mathbb{R}_{<a}$ and $\lambda \in S_{\delta, \phi}$.

- (i) Since $\lambda \in S_{\delta, \phi} \subset S_{a, \phi} \subset \rho(\mathbb{A})$, we immediately get that $(\lambda - \mathbb{A}) : D(\mathbb{A}) \rightarrow H$ is invertible. Now we observe that

$$|\lambda - a| \geq (a - \delta) \sin(\phi) > 0.$$

Using that, we get

$$\begin{aligned} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(H, H)} &\leq \frac{M_{\mathbb{A}}}{|\lambda - a|} = M_{\mathbb{A}} \frac{1 + |\lambda|}{(1 + |\lambda|)|\lambda - a|} \\ &\leq M_{\mathbb{A}} \frac{1 + a + |\lambda - a|}{|\lambda - a|} \frac{1}{1 + |\lambda|} \\ &\leq M_{\mathbb{A}} \frac{1 + a + (a - \delta) \sin(\phi)}{(a - \delta) \sin(\phi)} \frac{1}{1 + |\lambda|} = \frac{M_1}{1 + |\lambda|}. \end{aligned}$$

By theorem 3.5.1 and definition 3.5.1, $\mathbb{A}^{\frac{1}{2}} : V \rightarrow H$ is a bijection. Since per definition $\|\mathbb{A}^{\frac{1}{2}} v\|_H = \|v\|_V$ and $\|\mathbb{A}^{-\frac{1}{2}} h\|_V = \|h\|_H$ for all $v \in V, h \in H$, we obviously have $\mathbb{A}^{\frac{1}{2}} \in \mathcal{L}(V, H)$ and $\mathbb{A}^{-\frac{1}{2}} \in \mathcal{L}(H, V)$ with $\|\mathbb{A}^{\frac{1}{2}}\|_{\mathcal{L}(V, H)} = \|\mathbb{A}^{-\frac{1}{2}}\|_{\mathcal{L}(H, V)} = 1$. Thus we get

$$\begin{aligned} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(V, V)} &= \|\mathbb{A}^{-\frac{1}{2}} (\lambda - \mathbb{A})^{-1} \mathbb{A}^{\frac{1}{2}}\|_{\mathcal{L}(V, V)} \\ &\leq \|\mathbb{A}^{-\frac{1}{2}}\|_{\mathcal{L}(H, V)} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(H, H)} \|\mathbb{A}^{\frac{1}{2}}\|_{\mathcal{L}(V, H)} \leq \frac{M_1}{1 + |\lambda|}. \end{aligned}$$

- (ii) We begin by showing that $(\lambda - \mathbb{A}) : V \rightarrow V'$ is a bijection. Let $f \in V'$. Because of $V = H_{\frac{1}{2}}(\mathbb{A}) = H_{\frac{1}{2}}(\mathbb{A}^*)$ with equivalent norms, $(x, y) \mapsto (\mathbb{A}^{*\frac{1}{2}} x, \mathbb{A}^{*\frac{1}{2}} y)_H$ is a continuous and coercive sesquilinear form on V . Thus, there exists a unique $v_f \in V$ so that $\langle f, w \rangle = (\mathbb{A}^{*\frac{1}{2}} v_f, \mathbb{A}^{*\frac{1}{2}} w)_H$ for all $w \in V$. Now we set $v := \mathbb{A}^{\frac{1}{2}} [(\lambda - \mathbb{A})|_{D(\mathbb{A})}]^{-1} \mathbb{A}^{*\frac{1}{2}} v_f$. Because of

$$\mathbb{A} [(\lambda - \mathbb{A})|_{D(\mathbb{A})}]^{-1} \mathbb{A}^{*\frac{1}{2}} v_f = h \in H$$

for some $h \in H$, we have

$$v = \mathbb{A}^{\frac{1}{2}} [(\lambda - \mathbb{A})|_{D(\mathbb{A})}]^{-1} \mathbb{A}^{*\frac{1}{2}} v_f = \mathbb{A}^{-\frac{1}{2}} h \in V.$$

For all $w \in V$ we then get

$$\begin{aligned} \langle (\lambda - \mathbb{A}) v, w \rangle &= (\lambda v, w)_H - \left(\mathbb{A}^{\frac{1}{2}} v, \mathbb{A}^{*\frac{1}{2}} w \right)_H \\ &= \left(\lambda [(\lambda - \mathbb{A})|_{D(\mathbb{A})}]^{-1} \mathbb{A}^{*\frac{1}{2}} v_f, \mathbb{A}^{*\frac{1}{2}} w \right)_H - \left(\mathbb{A} [(\lambda - \mathbb{A})|_{D(\mathbb{A})}]^{-1} \mathbb{A}^{*\frac{1}{2}} v_f, \mathbb{A}^{*\frac{1}{2}} w \right)_H \\ &= \left((\lambda - \mathbb{A}) [(\lambda - \mathbb{A})|_{D(\mathbb{A})}]^{-1} \mathbb{A}^{*\frac{1}{2}} v_f, \mathbb{A}^{*\frac{1}{2}} w \right)_H \\ &= \left(\mathbb{A}^{*\frac{1}{2}} v_f, \mathbb{A}^{*\frac{1}{2}} w \right)_H = \langle f, w \rangle. \end{aligned}$$

Hence $\lambda - \mathbb{A}$ is surjective.

Now let $u \in V$ so that $\langle (\lambda - \mathbb{A})u, w \rangle = 0$ for all $w \in V$. By proposition 3.5.3 we have $\bar{\lambda} \in \rho(\mathbb{A}^*)$. So we can set $z := \mathbb{A}^{*\frac{1}{2}} [(\bar{\lambda} - \mathbb{A}^*)|_{D(\mathbb{A}^*)}]^{-1} \mathbb{A}^{\frac{1}{2}} u$. As in the surjectivity proof above, we see that $z \in V$ by observing that

$$\mathbb{A}^* [(\bar{\lambda} - \mathbb{A}^*)|_{D(\mathbb{A}^*)}]^{-1} \mathbb{A}^{\frac{1}{2}} u = g \in H$$

for some $g \in H$, so

$$\mathbb{A}^{*\frac{1}{2}} [(\bar{\lambda} - \mathbb{A}^*)|_{D(\mathbb{A}^*)}]^{-1} \mathbb{A}^{\frac{1}{2}} u = \mathbb{A}^{*- \frac{1}{2}} g \in V.$$

We then get

$$\begin{aligned} 0 &= \langle (\lambda - \mathbb{A})u, z \rangle = (\lambda u, z)_H - \left(\mathbb{A}^{\frac{1}{2}} u, \mathbb{A}^{*\frac{1}{2}} z \right)_H \\ &= \left(\mathbb{A}^{\frac{1}{2}} u, \bar{\lambda} [(\bar{\lambda} - \mathbb{A}^*)|_{D(\mathbb{A}^*)}]^{-1} \mathbb{A}^{\frac{1}{2}} u \right)_H - \left(\mathbb{A}^{\frac{1}{2}} u, \mathbb{A}^* [(\bar{\lambda} - \mathbb{A}^*)|_{D(\mathbb{A}^*)}]^{-1} \mathbb{A}^{\frac{1}{2}} u \right)_H \\ &= \left(\mathbb{A}^{\frac{1}{2}} u, (\bar{\lambda} - \mathbb{A}^*) [(\bar{\lambda} - \mathbb{A}^*)|_{D(\mathbb{A}^*)}]^{-1} \mathbb{A}^{\frac{1}{2}} u \right)_H \\ &= \left(\mathbb{A}^{\frac{1}{2}} u, \mathbb{A}^{\frac{1}{2}} u \right)_H = \|u\|_V^2. \end{aligned}$$

So $\lambda - \mathbb{A}$ is injective.

Using (ii) of proposition 3.6.1 above and the definition of $\lambda : V \rightarrow V'$ from this theorem, we see that $(\lambda - \mathbb{A}) \in \mathcal{L}(V, V')$. From the open mapping theorem we then get $(\lambda - \mathbb{A})^{-1} \in \mathcal{L}(V', V)$ as well.

Since we have $a > 0$, we can choose $\delta \in (\infty, a)$ so that $0 \in S_{\delta, \phi}$. Thus, the above proof shows in particular that $\mathbb{A} : V \rightarrow V'$ is invertible with $\mathbb{A} \in \mathcal{L}(V, V')$ and $\mathbb{A}^{-1} \in \mathcal{L}(V', V)$. Applying that and the second bound from part (i) of this theorem, we deduce

$$\begin{aligned} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(V', V')} &= \|\mathbb{A}(\lambda - \mathbb{A})^{-1} \mathbb{A}^{-1}\|_{\mathcal{L}(V', V')} \\ &\leq \|\mathbb{A}\|_{\mathcal{L}(V, V')} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(V, V)} \|\mathbb{A}^{-1}\|_{\mathcal{L}(V', V)} \leq \frac{M_2}{1 + |\lambda|} \end{aligned}$$

and

$$\begin{aligned} \|(\lambda - \mathbb{A})^{-1}\|_{\mathcal{L}(V', V)} &= \|(\lambda - \mathbb{A})^{-1} \mathbb{A} \mathbb{A}^{-1}\|_{\mathcal{L}(V', V)} \\ &\leq \|(\lambda - \mathbb{A})^{-1} \mathbb{A}\|_{\mathcal{L}(V, V)} \|\mathbb{A}^{-1}\|_{\mathcal{L}(V', V)} \\ &= \|\lambda(\lambda - \mathbb{A})^{-1} - 1\|_{\mathcal{L}(V, V)} \|\mathbb{A}^{-1}\|_{\mathcal{L}(V', V)} \\ &\leq |\lambda| \frac{M_2}{1 + |\lambda|} \|\mathbb{A}^{-1}\|_{\mathcal{L}(V', V)} + \|\mathbb{A}^{-1}\|_{\mathcal{L}(V', V)} \leq M_3. \end{aligned}$$

□

3.6.1 Examples of Extended Operators

Now we will apply the definitions and results from above to the Stokes and convection-diffusion operators. Again, as in sections 3.3 and 3.4, let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded Domain with

“sufficiently smooth” boundary and let all function spaces be *complex*-valued. We first observe the following:

Remark 3.6.1. Let \mathbb{S} be the Stokes operator defined as in section 3.4, then we have $\mathbb{S} = \mathbb{S}^*$ and thus obviously $H_{\frac{1}{2}}(\mathbb{S}) = H_{\frac{1}{2}}(\mathbb{S}^*)$ with equivalent norms. So \mathbb{S} extends to an operator on $H_{\frac{1}{2}}(\mathbb{S}) = V_\sigma(\Omega)$ — see section 3.5.1 for the last equality.

By basically the same arguments, the vector-valued Dirichlet Laplacian $\vec{\mathbb{L}}$ defined as in section 3.5.1 extends to an operator on $H_{\frac{1}{2}}(\vec{\mathbb{L}}) = H^{-1}(\Omega)^d$. In addition, $\vec{\mathbb{L}}$ and \mathbb{S} map elements of $V_\sigma(\Omega)$ to the same elements of $[V_\sigma(\Omega)]'$ when embedding $H^{-1}(\Omega)^d$ into $[V_\sigma(\Omega)]'$ in the usual manner.

Proof. We only show the last claim as the others are clear.

The orthogonality of the Leray projection and the definition of the Stokes operator provide the following for any $v \in D(\mathbb{S}) = V_\sigma(\Omega) \cap H^2(\Omega)^d$ and $w \in V_\sigma(\Omega)$:

$$\langle \vec{\mathbb{L}}v, w \rangle = (-\nabla v, w)_{L^2} = (-P_\sigma \nabla v, w)_{L^2} = \langle \mathbb{S}v, w \rangle.$$

Now, if we only have $v \in V_\sigma(\Omega)$, let $v_\epsilon \in D(\mathbb{S})$ with $\|v - v_\epsilon\|_{H^1} < \epsilon$ for any given $\epsilon > 0$. Then we get

$$\begin{aligned} |\langle \vec{\mathbb{L}}v, w \rangle - \langle \mathbb{S}v, w \rangle| &= |\langle \vec{\mathbb{L}}(v - v_\epsilon), w \rangle + \langle \vec{\mathbb{L}}v_\epsilon - \mathbb{S}v_\epsilon, w \rangle + \langle \mathbb{S}(v_\epsilon - v), w \rangle| \\ &\leq \|\vec{\mathbb{L}}\|_{\mathcal{L}(H_0^1, H^{-1})} \|v - v_\epsilon\|_{H^1} \|w\|_{H^1} \\ &\quad + \|\mathbb{S}\|_{\mathcal{L}(V_\sigma, V_\sigma')} \|v_\epsilon - v\|_{H^1} \|w\|_{H^1}, \end{aligned}$$

so $\langle \vec{\mathbb{L}}v, w \rangle = \langle \mathbb{S}v, w \rangle$ as well. □

For the convection-diffusion operator, we again have to be a bit more careful because of the asymmetry introduced by the convection. Nevertheless, we get the following without much difficulty.

Remark 3.6.2. Let \mathbb{D} be the convection-diffusion operator defined as in section 3.3 with the diffusion coefficient $\nu > 0$, the convection direction $b \in W^{1,\infty}(\Omega, \mathbb{R})^d$ and the reaction coefficient $c \in L^\infty(\Omega, \mathbb{R})$. Notice that unlike in section 3.3, we require the stronger regularity $b \in W^{1,\infty}(\Omega, \mathbb{R})^d$.

We then have $H_{\frac{1}{2}}(\mathbb{D}) = H_{\frac{1}{2}}(\mathbb{D}^*)$ with equivalent norms. So \mathbb{D} extends to an operator on $H_{\frac{1}{2}}(\mathbb{D}) = H_0^1(\Omega)$ — see again section 3.5.1 for the last equality.

Proof. We show $H_{\frac{1}{2}}(\mathbb{D}) = H_{\frac{1}{2}}(\mathbb{D}^*)$ with equivalent norms. The other claims then immediately follow using previous results.

From section 3.3 we know that \mathbb{D} is sectorial with sector $S_{a,\phi}$ for some $a \in \mathbb{R}$ and $\phi \in (0, \frac{\pi}{2})$. Let $\tilde{\mathbb{D}} := \mathbb{D} - a + \epsilon$ for some $\epsilon > \|\operatorname{div} b\|_{L^\infty}$. It is easy to see that $D(\tilde{\mathbb{D}}) \subset D(\tilde{\mathbb{D}}^*)$. Let $F(\tilde{\mathbb{D}})$ be the formal adjoint of $\tilde{\mathbb{D}}$, i.e.,

$$F(\tilde{\mathbb{D}})u := -\nu \Delta u - (b \cdot \nabla)u + (c - \operatorname{div} b - a + \epsilon)u,$$

for all $u \in D(\tilde{\mathbb{D}}) := D(\mathbb{D}) = H_0^1(\Omega) \cap H^2(\Omega)$.

$F(\tilde{\mathbb{D}})$ is just a convection-diffusion operator with different parameters, so the techniques from 3.3 can be used to show that $F(\tilde{\mathbb{D}})$ is a sectorial operator. The choice of ϵ makes it so that $F(\tilde{\mathbb{D}})$ is sectorial with a sector that is contained in the set $\{z \in \mathbb{C} : \operatorname{Re} z > 0\}$. In particular, $F(\tilde{\mathbb{D}})$ is invertible.

Because $F(\tilde{\mathbb{D}})$ is the formal adjoint of $\tilde{\mathbb{D}}$, it holds for all $x, y \in D(\tilde{\mathbb{D}})$ that $(\tilde{\mathbb{D}}x, y)_{L^2} = (x, F(\tilde{\mathbb{D}})y)_{L^2}$. So on $D(\tilde{\mathbb{D}})$, we have $\tilde{\mathbb{D}}^* = F(\tilde{\mathbb{D}})$ and since $F(\tilde{\mathbb{D}})$ is invertible, $\tilde{\mathbb{D}}^*$ is at least surjective.

Because of

$$\mathcal{N}(\tilde{\mathbb{D}}^*) = \mathcal{R}(\tilde{\mathbb{D}})^\perp = L^2(\Omega)^\perp = \{0\},$$

$\tilde{\mathbb{D}}^*$ is also injective — see B.4 in the appendix of [24] for the first equality.

Now if there was a $y \in L^2(\Omega) \setminus D(\tilde{\mathbb{D}})$ with $y \in D(\tilde{\mathbb{D}}^*)$, there would be an $x \in L^2(\Omega)$ and a $\tilde{y} \in D(\tilde{\mathbb{D}})$ with $\tilde{\mathbb{D}}^*y = x = F(\tilde{\mathbb{D}})\tilde{y} = \tilde{\mathbb{D}}\tilde{y}$, rendering $\tilde{\mathbb{D}}^*$ **not** injective. \nmid

Thus, we have $D(\tilde{\mathbb{D}}^*) \subset D(\tilde{\mathbb{D}})$ and consequently $D(\tilde{\mathbb{D}}^*) = D(\tilde{\mathbb{D}})$ as well as $F(\tilde{\mathbb{D}}) = \tilde{\mathbb{D}}^*$.

From remark 3.5.1 we then get

$$H_{\frac{1}{2}}(\mathbb{D}) = H_{\frac{1}{2}}(\tilde{\mathbb{D}}) = H_0^1(\Omega) = H_{\frac{1}{2}}(F(\tilde{\mathbb{D}})) = H_{\frac{1}{2}}(\tilde{\mathbb{D}}^*) = H_{\frac{1}{2}}(\mathbb{D}^* - a + \epsilon) = H_{\frac{1}{2}}(\mathbb{D}^*)$$

with norm equivalence. The second to last equality is easily seen with the definition 3.1.2 of the adjoint. \square

In sections 3.3 and 3.4 we constructed solutions by essentially using the Lax-Milgram theorem with sesquilinear forms that were associated with the operators. Now we see that the domains of those sesquilinear forms are exactly the domains of the square roots of the operators.

If for example $\mathbb{S} : D(\mathbb{S}) \rightarrow H_\sigma(\Omega)$, $D(\mathbb{S}) = V_\sigma(\Omega) \cap H^2(\Omega)^d$, is again the Stokes operator as it is defined in section 3.4 and $B_\mathbb{S}$ its associated sesquilinear form on $V_\sigma(\Omega) \times V_\sigma(\Omega)$, then the extension of \mathbb{S} to $V := H_{\frac{1}{2}}(\mathbb{S})$ — lets call it \mathbb{S}_V for clarity here — matches $B_\mathbb{S}$ exactly. I.e. we have $V = H_{\frac{1}{2}}(\mathbb{S}) = V_\sigma(\Omega)$ and for any $v \in V$ and $f \in V'$ we have $\mathbb{S}_V v = f$ if and only if $B_\mathbb{S}(v, \phi) = \langle f, \phi \rangle$ for all $\phi \in V$.

3.7 Problem Statement for Semilinear Parabolic Equations (SPEs)

We will now give a very general problem formulation. Once again, let H denote a separable complex Hilbert space.

Problem 3.7.1 (Initial Value Problem for the Semilinear Parabolic Equation (SPE)).

Let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a sectorial operator and let $V := H_{\frac{1}{2}}(\mathbb{A}) = H_{\frac{1}{2}}(\mathbb{A}^*)$ with equivalent norms so that, as in proposition 3.6.1, (V, H, V') forms a triplet of spaces and \mathbb{A} can be extended to a bounded operator $\mathbb{A} : V \rightarrow V'$.

Furthermore, let $T \in \mathbb{R}_{>0}$, let $G : (0, T] \rightarrow V'$ be the forcing term and let the possibly nonlinear term $N : V \rightarrow V'$ fulfill some kind of Lipschitz condition (see definition 3.7.1 or theorem 3.7.3 below). At last, let $u_0 \in H$ be the initial condition.

We then seek a $\tilde{T} \in (0, T]$ as large as possible and a function $u \in C([0, \tilde{T}]; H)$ with $u(t) \in V$, $\frac{du}{dt}(t) \in V'$ for all $t \in (0, \tilde{T}]$ and

$$\frac{du}{dt}(t) + \mathbb{A}u(t) = N(u(t)) + G(t) \quad \text{for all } t \in (0, \tilde{T}], \quad (3.12)$$

$$u(0) = u_0. \quad (3.13)$$

Even though we are working in abstract Hilbert spaces here — so the usual definition of *parabolic* PDEs does not apply — the condition that \mathbb{A} is a sectorial operator can be seen as a generalization of that classic definition. The requirement 3.7.1 for the nonlinearity can then be understood as a mathematically precise way to describe that the linear part dominates the equation (in PDEs one could say that the nonlinearity only contains lower order derivatives). In accordance with the literature (see for example the title of [27]), we call an equation with this kind of nonlinearity and solution-independent sectorial operator a *semilinear parabolic equation* (SPE).

A generalization are the so-called quasi-linear equations, which are allowed to have solution dependent operators and are extensively covered for example in [4].

Notice that our specific way of creating the triplet of spaces from fractional powers of operators is not standard. We have included it in the problem formulation here because it will always fit our applications and allows us to easily use results from section 3.6.

Of course, it cannot be shown in general that the above problem always has a unique solution on the whole interval $[0, T]$. The Navier-Stokes problem is a prime example for that. Nevertheless — depending on the regularity of the solution, the initial condition, the nonlinearity and the forcing term — at least *local* existence and uniqueness results have been shown for this type of problem by different authors.

Before we discuss some of those results, we will define the type of nonlinearity that our problems will often have.

Definition 3.7.1 (Locally Lipschitz Continuous). *Let $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ be a sectorial operator. In addition, let $\alpha \in [0, 1]$ and let X, Y be some sets with $H_\alpha(\mathbb{A}) \subset X$ and $H \subset Y$. Let the function $F : X \rightarrow Y$ map $H_\alpha(\mathbb{A})$ into H . We then call F locally Lipschitz continuous with respect to $H_\alpha(\mathbb{A})$ if there exists a constant $C_F > 0$ so that the following holds for all $x_1, x_2 \in H_\alpha(\mathbb{A})$:*

$$\|F(x_1) - F(x_2)\|_H \leq C_F (1 + \|x_1\|_{\mathbb{A}^\alpha} + \|x_2\|_{\mathbb{A}^\alpha}) \|x_1 - x_2\|_{\mathbb{A}^\alpha}.$$

Now, we give — without proofs — some results on the solvability of problem 3.7.1.

Theorem 3.7.1. *Augment problem 3.7.1 with the following requirements:*

- (a) *Let the forcing term G map $(0, T]$ into H and let it be locally Hölder continuous.*
- (b) *For some $\alpha \in [0, 1)$ let the nonlinearity N map $H_\alpha(\mathbb{A})$ into H and let it be locally Lipschitz continuous with respect to $H_\alpha(\mathbb{A})$ in the sense of 3.7.1.*

(c) Let the initial value have the regularity $u_0 \in H_\alpha(\mathbb{A})$.

Then there is a $\tilde{T} \leq T$ and a function $u : [0, \tilde{T}] \rightarrow H$ which (locally) solves the so augmented problem 3.7.1 and is unique with the following additional properties:

(i) The solution has the regularity $u \in C([0, \tilde{T}]; H)$ with $u(t) \in D(\mathbb{A})$ and $\frac{du}{dt}(t) \in H$ for all $t \in (0, \tilde{T}]$.

(ii) The mapping $[0, \tilde{T}] \ni t \mapsto N(u(t)) + G(t) \in H$ is locally Hölder continuous with

$$\int_0^\epsilon \|N(u(t)) + G(t)\|_H dt < \infty$$

for some $\epsilon \in (0, \tilde{T}]$.

(iii) The final time \tilde{T} depends on the initial value u_0 and if N and G are “suitably” (see corollary 3.3.5. in [27] for details) bounded, then the solution exists globally.

The above is an earlier result due to Henry — see theorem 3.3.3. and corollary 3.3.5. in [27].

From the more recent book [73] by Yagi we get the following result:

Theorem 3.7.2. *Augment 3.7.1 with the following requirements:*

(a) For some $\alpha \in (0, 1)$ let the nonlinearity N map $H_\alpha(\mathbb{A})$ into H and let it be locally Lipschitz continuous with respect to $H_\alpha(\mathbb{A})$ in the sense of 3.7.1.

(b) Let $\sigma \in (0, 1 - \alpha)$ and let the forcing term G be in the function space $\mathcal{F}^{\alpha, \sigma}((0, T]; H)$ of weighted Hölder continuous functions (see chapter 1 section 2.4 of [73] for details).

(c) Let the initial value have the regularity $u_0 \in H_\alpha(\mathbb{A})$.

Then there is a $\tilde{T} \leq T$ and a function $u : [0, \tilde{T}] \rightarrow H$ which (locally) solves the so augmented problem 3.7.1 and is unique with the following additional properties:

(i) The solution has the regularity $u \in C([0, \tilde{T}]; H_\alpha(\mathbb{A}))$ with the functions $(0, \tilde{T}] \ni t \mapsto u(t) \in H_1(\mathbb{A})$, $(0, \tilde{T}] \ni t \mapsto \frac{du}{dt}(t) \in H$ being continuous and with $\frac{du}{dt}, \mathbb{A}u \in \mathcal{F}^{\alpha, \sigma}((0, T]; H)$.

(ii) u fulfills the bound

$$\|u\|_{C([0, \tilde{T}]; H_\alpha(\mathbb{A}))} + \left\| \frac{du}{dt} \right\|_{\mathcal{F}^{\alpha, \sigma}((0, T]; H)} + \|\mathbb{A}u\|_{\mathcal{F}^{\alpha, \sigma}((0, T]; H)} < C_{G, u_0}$$

with some constant $C_{G, u_0} > 0$.

(iii) The final time \tilde{T} depends only on $\|G\|_{\mathcal{F}^{\alpha, \sigma}((0, T]; H)}$ and $\|u_0\|_{\mathbb{A}^\alpha}$.

If it is a priori known for possible solutions v of this type on larger intervals $[0, T_v]$, $T \geq T_v \geq T_{G, u_0}$ that they would always stay bounded by C_{G, u_0} as well, i.e.,

$$\|v\|_{C([0, T_v]; H_\alpha(\mathbb{A}))} < C_{G, u_0},$$

then the unique solution u already exists on the whole interval $[0, T]$.

We want to mention that for more regular initial data and a more regular forcing term — for example $u_0 \in H_\gamma(\mathbb{A})$ and $G \in \mathcal{F}^{\gamma,\sigma}((0, T]; H)$ for some $\gamma \in (\alpha, 1]$ — the solution u can be shown to also have more regularity, in particular we then have the stronger bound

$\|u\|_{C([0, \tilde{T}]; H_\gamma(\mathbb{A}))} < C_{G, u_0}$. The nonlinearity N and the operator \mathbb{A} do **not** need to fulfill stronger requirements. See theorem 4.2 in [73] for details.

In both of the above results, the solution u can be expressed using the semigroup notation (see theorem 3.2.1) in the following way:

$$u(t) = e^{-t\mathbb{A}}u_0 + \int_0^t e^{-(t-s)\mathbb{A}}[N(u(s)) + G(s)]ds \quad 0 \leq t \leq \tilde{T}.$$

Lastly, we want to list a result that stems from a variational view of the problem.

Theorem 3.7.3. *Augment 3.7.1 with the following requirements:*

(a) *Let B be a continuous and coercive sesquilinear form on V so that $\mathbb{A} : V \rightarrow V'$ is the operator corresponding to that sesquilinear form, i.e., $\langle \mathbb{A}u, v \rangle = B(u, v)$ for all $u, v \in V$.*

Furthermore, let the embedding $V \xrightarrow{d} H$ be compact. This is for example the case when the resolvent of $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ is compact (see theorem 1.4.8. in [27] for details).

(b) *Let $G \in L^2(0, T; V')$.*

(c) *Let the nonlinearity N fulfill the following requirement. For any $\xi > 0$, there exist continuous increasing functions $\phi_\xi, \psi_\xi : \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ so that*

$$\begin{aligned} \|N(u)\|_{V'} &\leq \xi \|u\|_V + \phi_\xi(\|u\|_H), \quad \text{and} \\ \|N(u) - N(v)\|_{V'} &\leq \xi \|u - v\|_V + (\|u\|_V + \|v\|_V) \psi_\xi(\|u\|_H + \|v\|_H + 1) \|u - v\|_H \end{aligned}$$

for all $u, v \in V$.

(d) *Let the initial value have the regularity $u_0 \in H$.*

Then there is a $\tilde{T} \leq T$ and a function $u : [0, \tilde{T}] \rightarrow H$ which (locally) solves the so augmented problem 3.7.1 and is unique with the following additional properties:

(i) *The solution has the regularity $u \in L^2(0, \tilde{T}; V) \cap C([0, \tilde{T}]; H) \cap H^1(0, \tilde{T}; V')$.*

(ii) *u admits to the bound*

$$\|u\|_{L^2(0, \tilde{T}; V)} + \|u\|_{C([0, \tilde{T}]; H)} + \|u\|_{H^1(0, \tilde{T}; V')} < C_{G, u_0}$$

with some constant $C_{G, u_0} > 0$.

(iii) *The final time \tilde{T} depends only on $\|G\|_{L^2(0, \tilde{T}; V')}$ and $\|u_0\|_H$.*

We want to quickly remark that in the linear case, i.e., $N \equiv 0$, the situation is naturally simpler. For example, in theorems 3.7.1 and 3.7.2, the initial value u_0 can be taken in the larger space H . See section 3.2 of [27] or chapter 3 section 2 of [73] for details on the linear case.

To close this section, we want to stress once again that the above existence and uniqueness results do not provide solutions for all problems that interest us — at least not on a whole given time interval. Again, the Navier-Stokes equations certainly fall into that category. Nevertheless, when presenting the numerical theory, we will always assume that a unique solution exists globally and has sufficient regularity.

3.8 The Incompressible Navier-Stokes Equations as an SPE

In this section, we will show how to reformulate the incompressible Navier-Stokes equations (2.13) and (2.14) into an initial value problem that fits into the framework we defined above in section 3.7.

We start by recalling a few previous definitions from section 3.2. Throughout this section let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded domain with “sufficiently smooth” boundary. We then remember that

$$\begin{aligned} V_\sigma(\Omega) &= \overline{\{v \in C_0^\infty(\Omega)^d : \operatorname{div} v = 0\}}^{\|\cdot\|_{H^1}} = \{v \in H_0^1(\Omega)^d : \operatorname{div} v = 0\}, \\ H_\sigma(\Omega) &= \overline{\{v \in C_0^\infty(\Omega)^d : \operatorname{div} v = 0\}}^{\|\cdot\|_{L^2}} = \{v \in L^2(\Omega)^d : \gamma_n(v) = 0, \operatorname{div} v = 0\}, \\ Q(\Omega) &= \{p \in L^2(\Omega) : (p, 1)_{L^2} = 0\}, \end{aligned}$$

with the first space being equipped with the $\|\cdot\|_{H^1}$ -norm, the last two spaces being equipped with the $\|\cdot\|_{L^2}$ -norm and the trace operator γ_n given in theorem 3.4.1. Furthermore, the Stokes operator \mathbb{S} was defined as

$$\mathbb{S} : D(\mathbb{S}) \subset H_\sigma(\Omega) \rightarrow H_\sigma(\Omega), v \mapsto -P_\sigma(\Delta v),$$

with $D(\mathbb{S}) := V_\sigma(\Omega) \cap H^2(\Omega)^d$ and the orthogonal Leray projection $P_\sigma : L^2(\Omega)^d \rightarrow H_\sigma(\Omega)$.

Remember that the Dirichlet Laplacian \mathbb{L} and the vector-valued Dirichlet Laplacian $\vec{\mathbb{L}}$ from section 3.5.1 were defined as

$$\begin{aligned} \mathbb{L} : D(\mathbb{L}) \subset L^2(\Omega) &\rightarrow L^2(\Omega), u \mapsto -\Delta u, \\ \vec{\mathbb{L}} : D(\vec{\mathbb{L}}) \subset L^2(\Omega)^d &\rightarrow L^2(\Omega)^d, u \mapsto -\Delta u \end{aligned}$$

with $D(\mathbb{L}) = H_0^1(\Omega) \cap H^2(\Omega)$ and $D(\vec{\mathbb{L}}) = H_0^1(\Omega)^d \cap H^2(\Omega)^d$.

We are now going to formulate an embedding result for the fractional spaces associated with \mathbb{L} , $\vec{\mathbb{L}}$ and \mathbb{S} . Parts of this result will be used to handle the nonlinearity in the Navier-Stokes problem.

Theorem 3.8.1.

(i) Let $d \in \{2, 3\}$, $q \in \mathbb{N}$ with $q \geq 2$ and $\alpha \in (\gamma, 1]$, where $\gamma := \frac{1}{2} + \frac{d}{4} \frac{q-2}{q}$. Then we have the following continuous embeddings:

$$H_\alpha(\mathbb{L}) \hookrightarrow W^{1,q}(\Omega), \quad H_\alpha(\vec{\mathbb{L}}) \hookrightarrow W^{1,q}(\Omega)^d, \quad H_\alpha(\mathbb{S}) \hookrightarrow W^{1,q}(\Omega)^d.$$

(ii) Let $d \in \{2, 3\}$ and $\alpha \in (\frac{d}{4}, 1]$. Then the following embeddings are well-defined and continuous.

$$H_\alpha(\mathbb{L}) \hookrightarrow C(\Omega), \quad H_\alpha(\vec{\mathbb{L}}) \hookrightarrow C(\Omega)^d, \quad H_\alpha(\mathbb{S}) \hookrightarrow C(\Omega)^d.$$

Intuitively, these results may be understood in two steps:

First, a more general version of proposition 3.5.4 provides a characterization of the fractional spaces associated with the operators as fractional Sobolev spaces.

Second, embedding theorems for fractional Sobolev spaces then provide the continuous embeddings into the Sobolev spaces and the spaces of continuous functions, respectively.

For a proof that avoids these steps — specifically one that does not use the characterizations of the fractional spaces associated with the operators as fractional Sobolev spaces — see section 1.6 and theorem 1.6.1. in [27]. We need one more important lemma before we can show our main result about the nonlinearity in the Navier-Stokes problem.

Lemma 3.8.1. *For any $v_1, v_2, v_3 \in H_0^1(\Omega)^d$ the following inequality holds:*

$$((v_1 \cdot \nabla) v_2, v_3)_{L^2} \leq C_N \|\nabla v_1\|_{L^2} \|\nabla v_2\|_{L^2} \|\nabla v_3\|_{L^2},$$

where $C_N > 0$ is a constant that only depends on Ω .

This can be proven with the Hölder inequality and some Sobolev embeddings from theorem 3.1.2. A detailed proof can be found for example in [8].

Let us now make the following definition:

Definition 3.8.1 (The Nonlinearity in Navier-Stokes). *For any $v \in H_0^1(\Omega)^d$ we define $N_S(v) \in L^1(\Omega)^d$ by setting*

$$N_S(v) := -(v \cdot \nabla)v.$$

We will promptly see that the mapping

$$H_0^1(\Omega)^d \ni w \mapsto (N_S(v), w)_{L^2}$$

is in $H^{-1}(\Omega)^d$. Hence, we will often identify $N_S(v)$ with that mapping and write $N_S(v) \in H^{-1}(\Omega)^d$.

We finally formulate the main result of this section:

Theorem 3.8.2.

(i) Let $v, w \in H_{\frac{1}{2}}(\vec{\mathbb{L}}) = H_0^1(\Omega)^d$ and let $C_N > 0$ be the constant from lemma 3.8.1. Then the following holds:

(a) The mapping $H_0^1(\Omega)^d \ni z \mapsto ((v \cdot \nabla) w, z)_{L^2}$ is in $H^{-1}(\Omega)^d$ with

$$|((v \cdot \nabla) w, z)_{L^2}| \leq C_N \|z\|_{H^1} \|v\|_{H^1} \|w\|_{H^1} \quad \text{for all } z \in H_0^1(\Omega)^d.$$

In particular we have $N_S(v) \in H^{-1}(\Omega)^d$.

(b)

$$\|N_S(v) - N_S(w)\|_{H^{-1}} \leq C_N (\|v\|_{H^1} + \|w\|_{H^1}) \|v - w\|_{H^1}.$$

(c) The Fréchet derivative $D^{(1)}N_S : H_0^1(\Omega) \rightarrow \mathcal{L}(H_0^1(\Omega)^d, H^{-1}(\Omega)^d)$ of N_S as a mapping from $H_0^1(\Omega)^d$ to $H^{-1}(\Omega)^d$, i.e., of the mapping $H_0^1(\Omega)^d \ni x \mapsto (N_S(x), \cdot)_{L^2} \in H^{-1}(\Omega)^d$, exists and fulfills

$$\|D^{(1)}N_S(v) - D^{(1)}N_S(w)\|_{\mathcal{L}(H_0^1, H^{-1})} \leq 2\|v - w\|_{H^1}.$$

Therefore the Fréchet derivative of N_S as a mapping from $V_\sigma(\Omega)$ to $[V_\sigma(\Omega)]'$, i.e., of the mapping $V_\sigma(\Omega) \ni x \mapsto (N_S(x), \cdot)_{L^2} \in [V_\sigma(\Omega)]'$, exists as well and fulfills an analogous bound in the appropriate norms.

(ii) Let $\alpha \in (\frac{3}{4}, 1]$ and $v, w \in H_0^1(\Omega)^d$ with $\{v, w\} \cap H_\alpha(\vec{\mathbb{L}}) \neq \emptyset$. Then the following holds:

(a) $(v \cdot \nabla) w \in L^2(\Omega)^d$ with

$$\begin{aligned} \|(v \cdot \nabla) w\|_{L^2} &\leq C_{\vec{\mathbb{L}}, 1} \|v\|_{\vec{\mathbb{L}}^\alpha} \|w\|_{H^1} && \text{if } v \in H_\alpha(\vec{\mathbb{L}}), \\ \|(v \cdot \nabla) w\|_{L^2} &\leq C_{\vec{\mathbb{L}}, 2} \|v\|_{H^1} \|w\|_{\vec{\mathbb{L}}^\alpha} && \text{if } w \in H_\alpha(\vec{\mathbb{L}}), \end{aligned}$$

where $C_{\vec{\mathbb{L}}, 1}, C_{\vec{\mathbb{L}}, 2} > 0$ are constants that depend on the embeddings given in theorem 3.8.1 and classical Sobolev embeddings. In particular we have $N_S(v) \in L^2(\Omega)^d$ and $P_\sigma[N_S(v)] \in H_\sigma(\Omega)$ for $v \in H_\alpha(\vec{\mathbb{L}})$.

(b) If both $v, w \in H_\alpha(\vec{\mathbb{L}})$, we have

$$\|N_S(v) - N_S(w)\|_{L^2} \leq C_{\vec{\mathbb{L}}} (\|v\|_{\vec{\mathbb{L}}^\alpha} + \|w\|_{\vec{\mathbb{L}}^\alpha}) \|v - w\|_{\vec{\mathbb{L}}^\alpha}.$$

Notice that for $\alpha \in [\frac{1}{2}, 1]$ we have the continuous embedding $H_\alpha(\mathbb{S}) \subset H_\alpha(\vec{\mathbb{L}})$. Thus, the above bounds in (a) and (b) also hold with $\|\cdot\|_{\mathbb{S}^\alpha}$ instead of $\|\cdot\|_{\vec{\mathbb{L}}^\alpha}$ on the right-hand side if v or w or both are in $H_\alpha(\mathbb{S})$, respectively.

Proof.

(i)

(a) This follows directly from lemma 3.8.1.

(b) Let $v, w, z \in H_{\frac{1}{2}}(\vec{\mathbb{L}}) = H_0^1(\Omega)^d$ and let $C_N > 0$ be the constant from lemma 3.8.1. Then we get

$$\begin{aligned} &|((v \cdot \nabla) v - (w \cdot \nabla) w), z)_{L^2}| \\ &= |((v \cdot \nabla) v - (w \cdot \nabla) v + (w \cdot \nabla) v - (w \cdot \nabla) w), z)_{L^2}| \\ &\leq |(((v - w) \cdot \nabla) v), z)_{L^2}| + |((w \cdot \nabla) (v - w)), z)_{L^2}| \\ &\leq C_N \|\nabla(v - w)\|_{L^2} \|\nabla v\|_{L^2} \|\nabla z\|_{L^2} + C_N \|\nabla w\|_{L^2} \|\nabla(v - w)\|_{L^2} \|\nabla z\|_{L^2}, \end{aligned}$$

which proves the claim.

(c) From (a) we get that for any $v \in H_0^1(\Omega)^d$ the mapping

$$H_0^1(\Omega)^d \ni w \mapsto -((v \cdot \nabla)w + (w \cdot \nabla)v, \cdot)_{L^2} \in H^{-1}(\Omega)^d$$

is in $\mathcal{L}(H_0^1(\Omega)^d, H^{-1}(\Omega)^d)$. It is easy to see that that mapping is the Fréchet derivative at v of N_S as a mapping from $H_0^1(\Omega)^d$ to $H^{-1}(\Omega)^d$. The estimate directly follows as well.

(ii)

(a) Let $\alpha \in (\frac{3}{4}, 1]$.

For $v \in H_\alpha(\vec{\mathbb{L}})$, $w \in H_0^1(\Omega)^d$ we get from (ii) of theorem 3.8.1

$$\|(v \cdot \nabla)w\|_{L^2} \leq \|v\|_{L^\infty} \|\nabla w\|_{L^2} \leq C_{\vec{\mathbb{L}}} \|v\|_{\vec{\mathbb{L}}^\alpha} \|w\|_{H^1}$$

with $C_{\vec{\mathbb{L}}} > 0$ depending on the embedding in (ii) of theorem 3.8.1.

For $v \in H_0^1(\Omega)^d$, $w \in H_\alpha(\vec{\mathbb{L}})$ we use (i) of theorem 3.8.1 and a generalized version of Hölder's inequality for the configuration $\frac{1}{2} = \frac{1}{6} + \frac{1}{3}$. We get

$$\begin{aligned} \|(v \cdot \nabla)w\|_{L^2} &\leq \|v\|_{L^6} \|\nabla w\|_{L^3} \leq \tilde{C}_{\vec{\mathbb{L}}} \|v\|_{L^6} \|w\|_{\vec{\mathbb{L}}^\alpha} \\ &\leq C_{\vec{\mathbb{L}}} \|v\|_{H^1} \|w\|_{\vec{\mathbb{L}}^\alpha} \end{aligned}$$

with $C_{\vec{\mathbb{L}}} > 0$ depending on the embedding in (ii) of theorem 3.8.1 and the classical Sobolev embedding $H^1(\Omega)^d \subset L^6(\Omega)^d$ from theorem 3.1.2.

(b) For $v, w \in H_\alpha(\vec{\mathbb{L}})$ this easily follows with $N_S(v) - N_S(w) = N_S(v) + (v \cdot \nabla)w - (v \cdot \nabla)w - N_S(w)$, the triangle inequality and the above part (a) of (ii).

Below, we will give two spatially weak formulations of the incompressible Navier-Stokes problem. One that is close to the original version as it was introduced in section 2.2 and a projected one that fits into our framework of SPEs. Our focus then lies on showing the equivalence of those two weak formulations — without delving deeper into the details of regularity, existence and uniqueness. For more details on weak formulations of the incompressible Navier-Stokes problem see for example chapter 3 of [68].

The weak formulation including a pressure is as follows:

Problem 3.8.1 (Incompressible Navier-Stokes Problem). *Let $T \in \mathbb{R}_{>0}$, $\nu \in \mathbb{R}_{>0}$, $v_0 \in H_\sigma(\Omega)$ and let $G(t) \in H^{-1}(\Omega)^d$ for all $t \in (0, T]$.*

Then we seek a pressure $p : (0, T] \rightarrow Q(\Omega)$ and a velocity $v \in C([0, T]; L^2(\Omega)^d)$ with $v(t) \in H_0^1(\Omega)^d$ and $\frac{dv}{dt}(t) = \frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t) \in H^{-1}(\Omega)^d$ for all $t \in (0, T]$ and

$$\frac{dv}{dt}(t) + \nu \vec{\mathbb{L}}v(t) + \text{grad } p(t) = N_S(v(t)) + G(t) \quad \text{for all } t \in (0, T], \quad (3.14)$$

$$\text{div } v(t) = 0 \quad \text{for all } t \in (0, T], \quad (3.15)$$

$$v(0) = v_0. \quad (3.16)$$

Notice that if $v(t) \in H_0^1(\Omega)^d$ for some $t \in [0, T]$, we have $N_S(v(t)) \in H^{-1}(\Omega)^d$ by theorem 3.8.2. Since the vector-valued Laplacian — see section 3.5.1 — extends to an operator from $H_0^1(\Omega)^d$

to $H^{-1}(\Omega)^d$ by the same arguments as we used for the Stokes operator, we also have $\vec{\mathbb{L}}v(t) \in H^{-1}(\Omega)^d$ for $v(t) \in H_0^1(\Omega)^d$. Furthermore, if $p(t) \in Q(\Omega)$ we have $\text{grad } p(t) \in V_\sigma^\perp(\Omega) \subset H^{-1}(\Omega)^d$ by part (i) of theorem 3.4.3. Thus, all terms in this formulation are well-defined.

Now we present the projected incompressible Navier-Stokes problem:

Problem 3.8.2 (Semilinear Parabolic Navier-Stokes Problem). *Let $T \in \mathbb{R}_{>0}$, $\nu \in \mathbb{R}_{>0}$, $v_0 \in H_\sigma(\Omega)$ and let $G(t) \in H^{-1}(\Omega)^d$ for all $t \in (0, T]$.*

Then we seek a velocity $v \in C([0, T]; H_\sigma(\Omega))$ with $v(t) \in V_\sigma(\Omega)$ and $\frac{dv}{dt}(t) = \frac{d[V_\sigma(\Omega) \ni x \mapsto (v, x)_{L^2}]}{dt}(t) \in [V_\sigma(\Omega)]'$ for all $t \in (0, T]$ and

$$\begin{aligned} \frac{dv}{dt}(t) + \nu \mathbb{S}v(t) &= N_S(v(t))|_{V_\sigma(\Omega)} + G(t)|_{V_\sigma(\Omega)} \quad \text{for all } t \in (0, T], \\ v(0) &= v_0. \end{aligned}$$

In the following remark, we gather properties of problem 3.8.2 which show how that problem fits into the framework of SPEs given in section 3.7:

Remark 3.8.1.

- (i) *The operator $\mathbb{S} : D(\mathbb{S}) \subset H_\sigma(\Omega) \rightarrow H_\sigma(\Omega)$ is sectorial, self-adjoint and invertible. In particular it extends to a bounded operator $\mathbb{S} : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$.*
- (ii) *For any $\alpha \in (\frac{3}{4}, 1]$, the nonlinearity N_S maps $H_\alpha(\vec{\mathbb{L}})$ into $L^2(\Omega)^d$ by (ii) of theorem 3.8.2. Thus, $P_\sigma N_S$ maps $H_\alpha(\mathbb{S})$ into $H_\sigma(\Omega)$ and is — again by (ii) of theorem 3.8.2 — locally Lipschitz continuous with respect to $H_\alpha(\mathbb{S})$ in the sense of definition 3.7.1. Notice that for any $v \in H_\alpha(\vec{\mathbb{L}}) \supset H_\alpha(\mathbb{S})$ we have $(P_\sigma N_S(v), w)_{L^2} = (N_S(v), w)_{L^2}$ for all $w \in V_\sigma(\Omega)$. Hence, $P_\sigma N_S(v)$ and $N_S(v)|_{V_\sigma(\Omega)}$ coincide as elements of $[V_\sigma(\Omega)]'$.*

Next, we want to show that the two weak formulations in problems 3.8.1 and 3.8.2 are largely equivalent. The main point here is, of course, to show the existence of a pressure in the non-projected formulation. As in the proof of remark 3.4.3, theorem 3.4.3 will be our tool for that.

Proposition 3.8.1. *Assume that the data in problems 3.8.1 and 3.8.2 coincides exactly — i.e., in both problems, the end time $T \in \mathbb{R}_{>0}$, the kinematic viscosity $\nu \in \mathbb{R}_{>0}$, the initial condition $v_0 \in H_\sigma(\Omega)$ and the forcing term $G : (0, T] \rightarrow H^{-1}(\Omega)^d$ are the same.*

- (i) *A velocity v then solves the projected SPE formulation 3.8.2 of the incompressible Navier-Stokes problem if and only if there exists a pressure p so that p and v solve the non-projected formulation 3.8.1 of the incompressible Navier-Stokes problem.*
- (ii) *Let $q \in \mathbb{N}_0$ and let v be a solution to problem 3.8.2 (so the velocity v and some pressure p solve problem 3.8.1). Then we have $v \in H^q(0, T; H_0^1(\Omega)^d)$ if and only if $v \in H^q(0, T; V_\sigma(\Omega))$.*

Proof.

- (i) \Leftarrow : Suppose that v and p solve problem 3.8.1, i.e., p is given as $p : (0, T] \rightarrow Q(\Omega)$ and we have $v \in C([0, T]; L^2(\Omega)^d)$ with $v(t) \in H_0^1(\Omega)^d$ and $\frac{dv}{dt}(t) = \frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t) \in H^{-1}(\Omega)^d$

for all $t \in (0, T]$.

Let $t \in (0, T]$. From (3.15) we immediately obtain $v(t) \in V_\sigma(\Omega) \subset H_\sigma(\Omega)$. Furthermore, for any $u \in V_\sigma(\Omega)$ we know that

$$\|[V_\sigma(\Omega) \ni x \mapsto (u, x)_{L^2}]\|_{[V_\sigma(\Omega)]'} \leq \|[H_0^1(\Omega)^d \ni x \mapsto (u, x)_{L^2}]\|_{H^{-1}}.$$

From that we can easily deduce

$$\frac{d[V_\sigma(\Omega) \ni x \mapsto (v, x)_{L^2}]}{dt}(t) = \frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t) \Big|_{V_\sigma(\Omega)} \in [V_\sigma(\Omega)]'.$$

Now let $w \in V_\sigma(\Omega)$. Then we observe that $\langle \text{grad } p(t), w \rangle = -(p(t), \text{div } w)_{L^2} = 0$ and by remark 3.6.1 also $\langle \vec{\mathbb{L}}v(t), w \rangle = \langle \mathbb{S}v(t), w \rangle$. Thus, we get:

$$\begin{aligned} & \left\langle \frac{d[V_\sigma(\Omega) \ni x \mapsto (v, x)_{L^2}]}{dt}(t), w \right\rangle + \nu \langle \mathbb{S}v(t), w \rangle - \langle N_S(v(t)), w \rangle \\ &= \left\langle \frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t), w \right\rangle + \nu \langle \vec{\mathbb{L}}v(t), w \rangle - \langle N_S(v(t)), w \rangle + \langle \text{grad } p(t), w \rangle \\ &= \langle G(t), w \rangle. \end{aligned}$$

Hence, v solves problem 3.8.2.

\Rightarrow : Let v solve problem 3.8.2, i.e., we have $v \in C([0, T]; H_\sigma(\Omega))$ with $v(t) \in V_\sigma(\Omega)$ and $\frac{dv}{dt}(t) = \frac{d[V_\sigma(\Omega) \ni x \mapsto (v, x)_{L^2}]}{dt}(t) \in [V_\sigma(\Omega)]'$ for all $t \in (0, T]$.

Let $t \in (0, T]$. Since $v(t) \in V_\sigma(\Omega)$, we immediately see that v fulfills (3.15). Moreover, some technical but not very difficult calculations, which involve the continuity of the Leray Projection P_σ provided by remark 3.4.2, show that

$$\frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t) = \frac{d[V_\sigma(\Omega) \ni x \mapsto (v, x)_{L^2}]}{dt}(t) \circ P_\sigma \in H^{-1}(\Omega)^d.$$

Next, we define

$$l(t) := -\frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t) - \nu \vec{\mathbb{L}}v(t) + N_S(v(t)) + G(t).$$

Since $\frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t) \in H^{-1}(\Omega)^d$ and $v(t) \in V_\sigma(\Omega) \subset H_0^1(\Omega)^d$, we have $l(t) \in H^{-1}(\Omega)$ by the observations under problem 3.8.1. And since v solves problem 3.8.2, we have for all $w \in V_\sigma(\Omega)$ again by remark 3.6.1:

$$\begin{aligned} \langle l(t), w \rangle &= -\left\langle \frac{d[H_0^1(\Omega)^d \ni x \mapsto (v, x)_{L^2}]}{dt}(t), w \right\rangle - \nu \langle \vec{\mathbb{L}}v(t), w \rangle + \langle N_S(v(t)), w \rangle + \langle G(t), w \rangle \\ &= -\left\langle \frac{d[V_\sigma(\Omega) \ni x \mapsto (v, x)_{L^2}]}{dt}(t), P_\sigma w \right\rangle - \nu \langle \mathbb{S}v(t), w \rangle + \langle N_S(v(t)), w \rangle + \langle G(t), w \rangle \\ &= 0, \end{aligned}$$

so $l(t) \in V_\sigma^\perp(\Omega)$. From part (i) of theorem 3.4.3 we now get that there exists a unique pressure $p(t) \in Q(\Omega)$ with $\text{grad } p(t) = l(t)$. Since t was arbitrary, p and v thus solve problem 3.8.1. \square

- (ii) If $v \in H^q(0, T; V_\sigma(\Omega))$, then $v \in H^q(0, T; H_0^1(\Omega)^d)$ follows directly from the definition of the weak (time) derivative in Bochner spaces — see definition 3.1.10.

If, on the other hand, we have $v \in H^q(0, T; H_0^1(\Omega)^d)$, then we know that $v(t) \in V_\sigma(\Omega)$ for all $t \in (0, T]$ because v solves problem 3.8.2. Now let $r \in \{0, \dots, q\}$. Since the Bochner integral is interchangeable with the application of a bounded linear operator, we obtain the following for all $\phi \in C_0^\infty(\Omega)$:

$$\begin{aligned} (-1)^r \int_0^T v^{(r)}(t) \phi(t) dt &= \int_0^T v(t) \frac{d^r \phi}{dt^r}(t) dt = \int_0^T P_\sigma(v(t)) \frac{d^r \phi}{dt^r}(t) dt \\ &= P_\sigma \left(\int_0^T v(t) \frac{d^r \phi}{dt^r}(t) dt \right) = P_\sigma \left((-1)^r \int_0^T v^{(r)}(t) \phi(t) dt \right) = (-1)^r \int_0^T P_\sigma(v^{(r)}(t)) \phi(t) dt. \end{aligned}$$

A variant of the fundamental lemma of calculus of variations then provides us with $P_\sigma(v^{(r)})(t) = v^{(r)}(t)$ for almost all $t \in (0, T)$, i.e., $v^{(r)}(t) \in V_\sigma(\Omega)$ for almost all $t \in (0, T)$. \square

3.8.1 The Oseen Operator

To close the chapter, we want to introduce an operator, which appears naturally as the linearization of the above discussed Navier-Stokes problem. When examining the application of some numerical methods to that problem, knowledge about the so-called Oseen operator will be very helpful.

Here we use all definitions and notations from the previous sections on the Stokes operator and the Navier-Stokes equations. And as in those sections, we need a bounded domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, with “sufficiently smooth” boundary, as well as the kinematic viscosity $\nu \in \mathbb{R}_{>0}$ and the end time $T \in \mathbb{R}_{>0}$ for problem 3.8.2. Also, let all function spaces be *complex*-valued. Furthermore, we assume the existence of a unique solution u with the regularity $u \in C([0, T]; H_\alpha(\mathbb{S}))$, $\alpha \in (\frac{3}{4}, 1]$, to that problem.

We now observe that for all $t \in [0, T]$, the Fréchet derivative at $u(t)$ of the function

$$V_\sigma(\Omega) \ni v \mapsto \nu \langle \mathbb{S}v, \cdot \rangle + ((v \cdot \nabla)v, \cdot)_{L^2} \in [V_\sigma(\Omega)]'$$

is given by

$$V_\sigma(\Omega) \ni v \mapsto \nu \langle \mathbb{S}v, \cdot \rangle + ((u(t) \cdot \nabla)v + (v \cdot \nabla)u(t), \cdot)_{L^2} \in [V_\sigma(\Omega)]'.$$

This prompts the following definition:

Definition 3.8.2 (Oseen Operator). *Let $t \in [0, T]$ and*

$$D(\mathbb{O}(t)) := D(\mathbb{S}) = V_\sigma(\Omega) \cap H^2(\Omega)^d$$

and let $\mathbb{O}(t) : D(\mathbb{O}(t)) \subset H_\sigma(\Omega) \rightarrow H_\sigma(\Omega)$ be the unbounded linear operator defined as

$$\mathbb{O}(t)v := \nu \mathbb{S}v + P_\sigma[(u(t) \cdot \nabla)v + (v \cdot \nabla)u(t)] \quad \text{for all } v \in D(\mathbb{O}(t)).$$

Then we call $\mathbb{O}(t)$ the Oseen operator at time t .

By (ii) of theorem 3.8.2, the operators are well-defined. Note that the definition of the operator $\mathbb{O}(t)$ depends on a function u that is omitted from the notation of the operator, i.e., we always assume that it is clear from the context, which function is used.

In the literature, the Oseen equation is usually defined in a way that does not exactly match our definition of the operator above. In the original equations for instance, there is no zero-order term for the velocity. In other definitions, the zero-order velocity coefficient is independent of the transport direction (whereas in our definition, $u(t)$ appears in both the convective and the zero-order term). We choose this way of defining the Oseen operator because then the operator matches exactly the above described Fréchet derivative — a fact that will be used later in the numerical theory.

Moreover, the zero-order term we use leads to a more interesting analysis. And while our definition does not technically *include* the other possible definitions for the Oseen operator, our results and proofs can be extended to most alternatives — essentially through simplification.

The operators $\mathbb{O}(t)$, $t \in [0, T]$, can be seen as perturbations of the Stokes operator. They are no longer symmetric, but their asymmetry only appears in lower order derivatives. In that sense the situation is similar to the relation between the convection-diffusion operator and the Dirichlet Laplacian.

The proof of the following result will use many concepts that were already shown in previous sections. Thus, we try to go into detail only where new ideas are used.

Theorem 3.8.3.

- (i) For each $t \in [0, T]$, the operator $\mathbb{O}(t)$ is densely-defined and closed.
- (ii) There exist $a \in \mathbb{R}$, $\phi \in (0, \frac{\pi}{2})$ and $M_{\mathbb{O}} > 0$, so that for each $t \in [0, T]$ and all $\lambda \in S_{a, \phi}$ we have

$$\|(\lambda - \mathbb{O}(t))^{-1}\|_{\mathcal{L}(H, H)} \leq \frac{M_{\mathbb{O}}}{|\lambda - a|},$$

i.e., the operators $\mathbb{O}(t)$, $t \in [0, T]$, are sectorial operators with constants that are independent of t .

- (iii) For each $t \in [0, T]$ and all $\alpha \in [\frac{1}{2}, 1]$, we have $H_{\alpha}(\mathbb{O}(t)) = H_{\alpha}(\mathbb{S})$ with norm equivalence **uniformly** in t , i.e., there are norm equivalence constants independent of t .

- (iv) For each $t \in [0, T]$, we have $H_{\frac{1}{2}}(\mathbb{O}(t)) = H_{\frac{1}{2}}(\mathbb{O}(t)^*) = V_{\sigma}(\Omega)$ with norm equivalence **uniformly** in t .

Proof. Throughout the whole proof, we will frequently use that for all $t \in [0, T]$ and all $v \in V_{\sigma}(\Omega)$ we have by (ii) (a) of theorem 3.8.2 that

$$\begin{aligned} \|(u(t) \cdot \nabla) v + (v \cdot \nabla) u(t)\|_{L^2} &\leq \tilde{C}(\Omega) \|u(t)\|_{\mathbb{S}^{\alpha}} \|v\|_{H^1} \\ &\leq \underbrace{\tilde{C}(\Omega) \|u\|_{C([0, T]; H_{\alpha}(\mathbb{S}))}}_{=: C(\Omega, u)} \|\nabla v\|_{L^2} \end{aligned} \tag{3.17}$$

with a constant $C(\Omega, u) > 0$ that does not depend on t .

- (i) Let $t \in [0, T]$. Then $D(\mathbb{O}(t))$ is certainly dense in $H_\sigma(\Omega)$ (see the beginning of the proof of theorem 3.4.5).

Now let $v \in D(\mathbb{O}(t)) = D(\mathbb{S})$, so $\mathbb{S}v \in H_\sigma(\Omega)$. Combining remark 3.4.3 and theorem 3.4.4 gives us

$$\|v\|_{H^2} \leq C_{SR} \|\mathbb{S}v\|_{L^2} \quad (3.18)$$

with some constant $C_{SR} > 0$ independent of v . Using (3.17), we get the estimate

$$\begin{aligned} \nu(\nabla v, \nabla v)_{L^2} &= \nu(\mathbb{S}v, v)_{L^2} = (\mathbb{O}(t)v - P_\sigma[(u(t) \cdot \nabla)v + (v \cdot \nabla)u(t)], v)_{L^2} \\ &\Rightarrow \nu\|\nabla v\|_{L^2}^2 \leq C_P(\Omega)\|\mathbb{O}(t)v\|_{L^2} + C(\Omega, u)\|v\|_{L^2} \end{aligned} \quad (3.19)$$

where $C_P(\Omega) > 0$ is the Poincaré constant.

With (3.18), (3.19) and again (3.17) we get

$$\|v\|_{H^2} \leq \left(C_{SR} + \frac{C_{SR}C(\Omega, u)C_P(\Omega)}{\nu} \right) \|\mathbb{O}(t)v\|_{L^2} + \frac{C_{SR}C(\Omega, u)^2}{\nu} \|v\|_{L^2}. \quad (3.20)$$

Proceeding as in the proof of proposition 3.3.1 then shows that $\mathbb{O}(t)$ is closed.

- (ii) Here we largely adapt the proof of theorem 3.3.3 in an abbreviated form.
Let $t \in [0, T]$ and define the sesquilinear form $B_{\mathbb{O}(t)}$ associated with $\mathbb{O}(t)$ by setting

$$B_{\mathbb{O}(t)}(v, w) := \nu(\nabla v, \nabla w)_{L^2} + ((u(t) \cdot \nabla)v, w)_{L^2} + ((v \cdot \nabla)u(t), w)_{L^2}$$

for all $v, w \in V_\sigma(\Omega)$.

Very similar to the way in which we got the Garding inequality in lemma 3.3.1 and using once again (3.17) we obtain for all $v \in V_\sigma(\Omega)$:

$$\operatorname{Re} B_{\mathbb{O}(t)}(v, v) \geq \frac{\nu}{2} \|\nabla v\|_{L^2}^2 - C_{G\mathbb{O}} \|v\|_{L^2}^2 \quad (3.21)$$

with $C_{G\mathbb{O}} := \frac{C(\Omega, u)^2}{2\nu}$. Analogously to lemma 3.3.2, one can then show — again with the help of (3.17) — that the numerical range

$$\Theta(B_{\mathbb{O}(t)}) := \{B_{\mathbb{O}(t)}(v, v) \in \mathbb{C} : v \in V_\sigma(\Omega), \|v\|_{L^2} = 1\}$$

of $B_{\mathbb{O}(t)}$ is contained in the set

$$U_{\mathbb{O}} := \left\{ \mu \in \mathbb{C} : \operatorname{Re} \mu \geq \frac{\nu}{2(C_P(\Omega))^2} - C_{G\mathbb{O}}, |\operatorname{Im} \mu| \leq C(\Omega, u) \sqrt{\frac{2}{\nu} (\operatorname{Re} \mu + C_{G\mathbb{O}})} \right\}.$$

Notice that $U_{\mathbb{O}}$ does not depend on t .

It is not hard to see that we can choose

$$a \in \mathbb{R}_{\leq -C_{G\mathbb{O}}} \quad \text{and} \quad \phi \in \left(0, \frac{\pi}{2}\right) \quad (3.22)$$

so that $S_{a,\phi} \subset \mathbb{C} \setminus \overline{U_\mathbb{O}}$. See the picture in the proof of theorem 3.3.3 for illustration.

Now choose any $\lambda \in S_{a,\phi}$. Using a version of the Lax-Milgram theorem for sesquilinear forms, we get that for any $f \in H_\sigma(\Omega)$ there exists a unique $v_f \in V_\sigma(\Omega)$ so that for all $w \in V_\sigma(\Omega)$ we have

$$\begin{aligned} \lambda(v_f, w)_{L^2} - B_{\mathbb{O}(t)}(v_f, w) &= (f, w)_{L^2} \\ \Leftrightarrow (\nabla v_f, \nabla w)_{L^2} &= \underbrace{\frac{\lambda}{\nu}(v_f, w)_{L^2} - \frac{1}{\nu}(f + (u(t) \cdot \nabla)v_f + (v_f \cdot \nabla)u(t), w)_{L^2}}_{=:(\tilde{f}, w)_{L^2}}. \end{aligned}$$

The Lax-Milgram theorem also provides us with a regularity estimate

$$\|v_f\|_{H^1} \leq C_\lambda \|f\|_{L^2} \quad (3.23)$$

where the constant $C_\lambda > 0$ only depends on λ , ν , $C_{G_\mathbb{O}}$ and $U_\mathbb{O}$.

Using again (3.17), we obtain $\tilde{f} \in L^2(\Omega)^d$ and thus, by combining remark 3.4.3 and theorem 3.4.4 as in the proof of (i), we get the higher regularity $v_f \in H^2(\Omega)^d$, i.e., $v_f \in D(\mathbb{O}(t))$.

Integration by parts quickly shows $(\lambda v_f - \mathbb{O}(t)v_f, w)_{L^2} = \lambda(v_f, w)_{L^2} - B_{\mathbb{O}(t)}(v_f, w) = (f, w)_{L^2}$ for all $w \in V_\sigma(\Omega)$. Thus, $\lambda - \mathbb{O}(t) : D(\mathbb{O}(t)) \rightarrow H_\sigma(\Omega)$ is a bijection. Furthermore, the bound (3.23) shows that $(\lambda - \mathbb{O}(t))^{-1} \in \mathcal{L}(H_\sigma(\Omega), H_\sigma(\Omega))$, i.e., $\lambda \in \rho(\mathbb{O}(t))$.

Using (i) of proposition 3.2.1, we finally get

$$\|(\lambda - \mathbb{O}(t))^{-1}\|_{\mathcal{L}(L^2, L^2)} \leq \frac{1}{\text{dist}(\lambda, \overline{\Theta(\mathbb{O}(t))})} \leq \frac{1}{\text{dist}(\lambda, \overline{U_\mathbb{O}})} \leq \frac{M_\mathbb{O}}{|\lambda - a|}$$

with the constant $M_\mathbb{O} > 0$ depending only on a , ϕ and $U_\mathbb{O}$. For a more detailed explanation of the last inequality we refer to the end of the proof of theorem 3.3.3 and the associated illustrative picture.

- (iii) Choose any $\epsilon > 0$ and let $t \in [0, T]$. With $a \in \mathbb{R}$ chosen as in part (ii) of this theorem and $\mathbb{A}_1 := \nu\mathbb{S}$, $\mathbb{A}_2 := \mathbb{O}(t) + \epsilon - a$, $\beta := \frac{1}{2}$ in theorem 3.5.2, it is not hard to prove our claim with the help of once again (3.17) and arguments similar to the ones that are used in the proof of remark 3.5.1.

The norm equivalence constant given in theorem 3.5.2 depends in our case only on $M_\mathbb{O}$, on a, ϕ from part (ii) of this theorem, on $C(\Omega, u)$ from (3.17), on the constant $C_{G_\mathbb{O}}$ from (3.21) and on the norm equivalence in part (iii) of proposition 3.5.4. Thus, we have norm equivalence **uniformly** in t .

- (iv) For all $t \in [0, T]$ and all $c \in \mathbb{R}$, the formal adjoint $F(\mathbb{O}(t) + c) : D(\mathbb{O}(t)) \subset H_\sigma(\Omega) \rightarrow H_\sigma(\Omega)$ of $(\mathbb{O}(t) + c) : D(\mathbb{O}(t)) \subset H_\sigma(\Omega) \rightarrow H_\sigma(\Omega)$ is given by

$$F(\mathbb{O}(t) + c)v := \nu\mathbb{S}v - (\overline{u}(t) \cdot \nabla)v + Y(\nabla u(t))v + cv$$

for all $v \in D(\mathbb{O}(t))$. Here, the matrix $Y(u(t)) \in (L^2(\Omega))^{d \times d}$ has the entries

$$[Y(u(t))]_{ij} := \frac{\partial \overline{u}_j(t)}{\partial x_i} \quad \text{for all } i, j \in \{1, \dots, d\}.$$

This expression for the formal adjoint can easily be deduced with integration by parts. Using again (3.17), it is clear that for all $v, w \in V_\sigma(\Omega)$ we have

$$|(-(\bar{u}(t) \cdot \nabla)v + Y(\nabla u(t))v, w)_{L^2}| \leq C_2(\Omega, u)\|v\|_{H^1}\|w\|_{L^2}$$

with a constant $C_2(\Omega, u) > 0$ that does not depend on t .

Analogously to the proof of part (i) and (ii) of this theorem, it can then be shown that $F(\mathbb{O}(t) + c)$ is a sectorial operator on $H_\sigma(\Omega)$. It is not hard to see that there exists a $c \in \mathbb{R}$ large enough, so that for all $t \in [0, T]$, both $F(\mathbb{O}(t) + c)$ and $\mathbb{O}(t) + c$ are sectorial with sectors that contain 0, i.e., the operators are invertible.

Our claim can then be validated by using part (iii) of the current theorem and techniques from the proof of remark 3.6.2. Since c can be chosen independent of t , and because the norm equivalences in part (iii) of the current theorem are also independent of t , the uniformity in t is once again verified.

□

The above theorem shows us that the operators $\mathbb{O}(t)$, $t \in [0, T]$, really are just asymmetric (and largely time-independent) perturbations of the Stokes operator — just as we claimed at the beginning of this section.

Now we want to apply proposition 3.6.1 to extend these operators to bounded operators — with all significant constants staying time-independent. We will also see that these extensions obtained from proposition 3.6.1 coincide exactly with a Fréchet derivative in the semilinear parabolic Navier-Stokes equation.

Theorem 3.8.4. *Let a , ϕ and $M_\mathbb{O}$ as in part (ii) of the theorem above. Then we have the following:*

(i) *For all $c \in \mathbb{R}$ and all $t \in [0, T]$, the operator*

$$(\mathbb{O}(t) + c) : D(\mathbb{O}(t)) \subset H_\sigma(\Omega) \rightarrow H_\sigma(\Omega)$$

is a sectorial operator on H with sector $S_{a+c, \phi}$ and sectoriality constant $M_\mathbb{O}$.

(ii) *For all $c \in \mathbb{R}$ and all $t \in [0, T]$, the (well-defined) extension of $\mathbb{O}(t) + c$ to a bounded operator from $V_\sigma(\Omega)$ to $[V_\sigma(\Omega)]'$ defined in (ii) of proposition 3.6.1 is given by*

$$\langle (\mathbb{O}(t) + c)v, w \rangle = \langle \nu \mathbb{S}v, w \rangle + ((u(t) \cdot \nabla)v + (v \cdot \nabla)u(t) + cv, w)_{L^2} \quad \text{for all } v, w \in V_\sigma(\Omega).$$

Thus, that extension is exactly the Fréchet derivative at $u(t)$ of the function

$$V_\sigma(\Omega) \ni v \mapsto \nu \langle \mathbb{S}v, \cdot \rangle + ((u(t) \cdot \nabla)v + (v \cdot \nabla)u(t) + cw, \cdot)_{L^2} \in [V_\sigma(\Omega)]'.$$

For all $c \in \mathbb{R}$ there exists a constant $C_\mathbb{O}$ so that for all $t \in [0, T]$ we have

$$\|\mathbb{O}(t) + c\|_{\mathcal{L}(V_\sigma, V_\sigma')} \leq C_\mathbb{O},$$

*i.e., the operators $(\mathbb{O}(t) + c) : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$, $t \in [0, T]$, are bounded **uniformly** in t .*

- (iii) For all $\epsilon \in \mathbb{R}_{>0}$ and all $t \in [0, T]$ the operator $(\mathbb{O}(t) + \epsilon - a) : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$ is invertible. Furthermore, for all $\epsilon \in \mathbb{R}$ there exists a constant $C_{\mathbb{O}, inv} > 0$ so that for all $t \in [0, T]$ we have

$$\|(\mathbb{O}(t) + \epsilon - a)^{-1}\|_{\mathcal{L}(V'_\sigma, V_\sigma)} \leq C_{\mathbb{O}, inv},$$

i.e., the inverses of the operators $(\mathbb{O}(t) + \epsilon - a) : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$, $t \in [0, T]$, are bounded **uniformly** in t .

- (iv) For all $\epsilon \in \mathbb{R}_{>0}$ the operators $\{(\lambda - (\mathbb{O}(t) + \epsilon - a)) : t \in [0, T], \lambda \in S_{\epsilon, \phi}\}$ are invertible and fulfill the strengthened resolvent bounds from theorem 3.6.1 **uniformly** in t . In particular, for all $\epsilon \in \mathbb{R}$ there exist constants $M_{\mathbb{O}, 1}, M_{\mathbb{O}, 2}, M_{\mathbb{O}, 3} > 0$, so that for all $t \in [0, T]$ and all $\lambda \in S_{\frac{\epsilon}{2}, \phi}$ we have

$$\begin{aligned} \|(\lambda - (\mathbb{O}(t) + \epsilon - a))^{-1}\|_{\mathcal{L}(H_\sigma, H_\sigma)} &\leq \frac{M_{\mathbb{O}, 1}}{1 + |\lambda|}, \\ \|(\lambda - (\mathbb{O}(t) + \epsilon - a))^{-1}\|_{\mathcal{L}(V_\sigma, V_\sigma)} &\leq \frac{M_{\mathbb{O}, 1}}{1 + |\lambda|}, \\ \|(\lambda - (\mathbb{O}(t) + \epsilon - a))^{-1}\|_{\mathcal{L}(V'_\sigma, V'_\sigma)} &\leq \frac{M_{\mathbb{O}, 2}}{1 + |\lambda|}, \\ \|(\lambda - (\mathbb{O}(t) + \epsilon - a))^{-1}\|_{\mathcal{L}(V'_\sigma, V_\sigma)} &\leq M_{\mathbb{O}, 3}. \end{aligned}$$

Proof.

- (i) We have often before used claims similar to this one — for example in the discussion under proposition 3.5.1 or in the proof of (iv) of the above theorem 3.8.3. These claims can easily be shown by a simple shift of the spectrum. We omit the details in this case as well.
- (ii) Let $t \in [0, T]$, $c \in \mathbb{R}$, $\epsilon > 0$ and $\tilde{\mathbb{O}}(t) := \mathbb{O}(t) + c - a - c + \epsilon = \mathbb{O}(t) - a + \epsilon$. By (iv) of theorem 3.8.3, we can apply (ii) of proposition 3.6.1 to obtain for all $v, w \in V_\sigma(\Omega) = H_{\frac{1}{2}}(\mathbb{O}(t) + c)$ the bounded extension

$$\langle (\mathbb{O}(t) + c)v, w \rangle := \left(\tilde{\mathbb{O}}(t)^{\frac{1}{2}}v, \tilde{\mathbb{O}}(t)^{* \frac{1}{2}}w \right)_{L^2} + (a + c - \epsilon)(v, w)_{L^2}.$$

For $w \in V_\sigma(\Omega)$ and the higher regularity $v \in D(\mathbb{O}(t))$, we get

$$\begin{aligned} \langle (\mathbb{O}(t) + c)v, w \rangle &= (\tilde{\mathbb{O}}(t)v, w)_{L^2} + (a + c - \epsilon)(v, w)_{L^2} \\ &= (\nu \mathbb{S}v + P_\sigma[(u(t) \cdot \nabla)v + (v \cdot \nabla)u(t)] + (\epsilon - a)v, w)_{L^2} \\ &\quad + (a + c - \epsilon)(v, w)_{L^2} \\ &= (\nu \mathbb{S}v + P_\sigma[(u(t) \cdot \nabla)v + (v \cdot \nabla)u(t)], w)_{L^2} + c(v, w)_{L^2} \\ &= (\nu \mathbb{S}v + (u(t) \cdot \nabla)v + (v \cdot \nabla)u(t), w)_{L^2} + c(v, w)_{L^2} \\ &= \langle \nu \mathbb{S}v, w \rangle + ((u(t) \cdot \nabla)v + (v \cdot \nabla)u(t) + cv, w)_{L^2}. \end{aligned}$$

Using again (3.17) further shows that for all $v, w \in V_\sigma(\Omega)$ we have

$$|\langle \nu \mathbb{S}v, w \rangle + ((u(t) \cdot \nabla)v + (v \cdot \nabla)u(t) + cv, w)_{L^2}| \leq (\nu + C(\Omega, u) + c) \|v\|_{H^1} \|w\|_{H^1}. \quad (3.24)$$

Thus the mapping $F : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$, defined by

$$\langle F(v), w \rangle := \langle \nu \mathbb{S}v, w \rangle + ((u(t) \cdot \nabla) v + (v \cdot \nabla) u(t) + cv, w)_{L^2} \quad \text{for all } v, w \in V_\sigma(\Omega),$$

fulfills $F \in \mathcal{L}(V_\sigma(\Omega), [V_\sigma(\Omega)]')$ and coincides with the extension $(\mathbb{O}(t) + c) \in \mathcal{L}(V_\sigma(\Omega), [V_\sigma(\Omega)]')$ — as it is given in (ii) of proposition 3.6.1 — on the set $D(\mathbb{O}(t))$, which is dense in $V_\sigma(\Omega)$. Hence the two mappings are equal on the whole of $V_\sigma(\Omega)$ and are bounded by the time-independent constant given in (3.24).

- (iii) Let $\epsilon > 0$ and $t \in [0, T]$. Using the expression for $(\mathbb{O}(t) + \epsilon - a) : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$ given in part (ii) of this theorem and revisiting the proof of theorem 3.8.3 part (ii) — mainly combining (3.22), (3.21) and a version of the Lax-Milgram theorem for sesquilinear forms — shows us that $(\mathbb{O}(t) - a + \epsilon) : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$ is invertible and the inverse is bounded.

The operator norm of the inverse is bounded by the constant given in the Lax-Milgram theorem — in theorem 3.3.1 and its proof, techniques to derive that constant in detail are shown for a different but similar operator. Closer inspection shows that in this case here, the constant only depends on $\epsilon, \nu, C_{G\mathbb{O}}$ and $U_{C\mathbb{O}}$. Thus it is independent of t .

- (iv) Part (iv) of theorem 3.8.3 lets us apply theorem 3.6.1 to obtain all parts of our claim here other than the time-independence of the constants.

From part (ii) and (iii) of the current theorem 3.8.4 we know that the operators $(\mathbb{O}(t) + \epsilon - a) : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$, $t \in [0, T]$, and their inverses can be bounded uniformly in t . In theorem 3.6.1 the constants for the resolvent bounds can be chosen to depend only on $\epsilon, a, \phi, M_{\mathbb{O}}$ and upper bounds for the operator norms of the operators $(\mathbb{O}(t) + \epsilon - a) : V_\sigma(\Omega) \rightarrow [V_\sigma(\Omega)]'$, $t \in [0, T]$, and their inverses. Thus, $M_{\mathbb{O},1}, M_{\mathbb{O},2}, M_{\mathbb{O},3} > 0$ can be chosen independent of t .

□

4 Discretization in Time by Rosenbrock-Type Methods

In this chapter we will introduce the numerical theory that we use to obtain a temporal discretization of semilinear parabolic equations. While there are many different ways to discretize parabolic equations in time, the focus of this work lies in examining the so-called W-methods and the related ROW methods. These methods were chosen specifically with the Navier-Stokes equations in mind for which they promised to be very useful for different reasons which will be explained throughout this chapter.

ROW methods and W-methods are among the Linearly-Implicit Runge Kutta-methods (LIRK) whose usage for semilinear parabolic problems is motivated by two facts:

- (i) The temporal discretization of parabolic problems very often requires highly stable — and thus **implicit** — schemes in order to circumvent having to use impractically small time steps in the resulting fully-discrete algorithms.
- (ii) Implicit schemes applied to nonlinear problems lead to nonlinear systems having to be solved — and that is often numerically expensive.

Roughly speaking, LIRK methods are designed in such a way that linearizations of nonlinear equations are handled implicitly and the — still nonlinear — error terms are treated explicitly. Thus, these methods are suitable for the time discretization of semilinear parabolic problems **if** the implicit handling of just the linearization is sufficient to obtain stable schemes. Since in semilinear parabolic equations a linear part essentially dominates the possibly nonlinear term (see the discussion at the beginning of section 3.7), it is reasonable to hope that LIRK methods are well-suited for these problems and will lead to stable schemes.

Out of the large class of LIRK methods, we will restrict ourselves to Rosenbrock-type methods which include the above mentioned ROW methods and W-methods. Other notable LIRK methods are for example the adaptive Runge-Kutta methods.

Before we begin with the time-stepping schemes however, we want to talk briefly about different concepts of organizing temporal and spatial discretization.

4.1 Rothe's Method

First of all, our SPEs are in a sense just ODEs with values in abstract Banach/Hilbert spaces, thus they can always be discretized in time. On the other hand, it is not clear what the notion of 'spatial discretization' would mean for those abstract spaces. One could understand spatial discretization — even for abstract Hilbert spaces — as looking for approximate solutions in finite dimensional subspaces, which is exactly what we will do later in chapter 5. Nevertheless, since the motivation for this work lies entirely in the handling of PDEs, the Banach/Hilbert spaces in question will always be Sobolev function spaces on spatial domains — leading to the usual, more tangible concepts of spatial discretization, such as the finite element method.

Our focus clearly lies on the time discretization, but for the sake of completeness we give at least an introduction into the spatial discretization as well — see chapter 5.

Our fully-discrete algorithm will be assembled by first discretizing in time our SPE, so that we obtain a sequence of Hilbert space equations, in which the formulation of each individual equation depends on the *fully*-discrete solutions of previous equations. In the usual Sobolev spaces that our applications will always be set in, this just means that we require the solution of certain stationary PDEs at each discrete time step. These stationary equations will then be discretized in space and solved by suitable methods like the finite element method to obtain the fully-discrete solution for that time-step.

This ordering of the discretization is often called Rothe's Method (see [57] for the original publication) and stands in contrast with the so-called Method of Lines (MOL) where the problems are first discretized in space, leading to large systems of classic ordinary differential equations (ODE), which are then solved with suitable ODE solvers. While the MOL was the more popular approach for quite a while because it allowed for easy usage of existing ODE solvers, nowadays Rothe's Method has also gained popularity because it can nicely incorporate the changing of spatial meshes between time steps and is thus well suited for adaptive refinement.

An additional advantage of discretizing in this order is that it is well suited to isolate the effects of temporal discretization by only doing one discretization step — i.e., semi-discretizing in time — and then analyzing the error without considering the spatial discretization. While this might not directly help in practical applications (where we usually need estimates on the fully-discrete error), it allows to examine how well a temporal discretization method will function if the mesh size becomes smaller and thus temporal stiffness increases — with stiffness essentially being infinite in the true semi-discretization.

Another approach that is different from the ones we talked about above is to employ a Galerkin procedure in space **and** time to obtain a method that is formulated in a space-time domain but can then be decoupled into a sequence of systems which correspond to individual time steps (see [55] for more details on that).

Throughout the whole chapter, we will assume the following scenario:

For the temporal discretization of a given SPE problem of the type 3.7.1, $M + 1 \in \mathbb{N}$ discrete points

$$0 = t_0 < t_1 < t_2 < \dots < t_{M-1} < t_M \leq T$$

are taken from the time interval $[0, T]$, $T \in \mathbb{R}_{>0}$, with the corresponding time step sizes $\tau_m := t_{m+1} - t_m$, $m \in \{0, \dots, M-1\}$. A solution $u : [0, T] \rightarrow H$ to the problem, with values in some Hilbert space H , is then approximated in these discrete points by a suitable numerical method, i.e., $u_m \approx u(t_m)$ for all $m \in \{0, \dots, M-1\}$.

In the following sections, instead of directly starting with the discretization in time of SPEs, we will take a detour to the field of ordinary differential equations. We do this because, as already mentioned, SPEs can be seen as ODEs with values in Hilbert spaces. So methods for the discretization of SPEs are really just methods for the discretization of ODEs where vectors from finite dimensional spaces have been replaced with Hilbert space objects. Hence, most properties of discretization methods for SPEs can be — and *are* in the literature — described with standard ODE terminology, such as convergence order, order conditions, stability function and others. This allows us to work in the simpler and easier to grasp setting of ODEs when introducing Rosenbrock-type methods and conveying the idea behind them.

4.2 Rosenbrock-Type Methods for Ordinary Differential Equations

In his paper [15], Deuffhard noted that when solving nonlinear ODEs with implicit integration schemes, just doing one Newton iteration with either the exact Jacobian or an approximation thereof could be enough to obtain stability. The Rosenbrock-type methods (see [56] for Rosenbrock's initial idea) use a very similar approach with the main difference being that the Jacobian is worked directly into the formula.

We will provide two ways of understanding the construction of Rosenbrock-type methods. Those ideas will be illustrated for ordinary differential equations of the following type:

Problem 4.2.1 (Initial Value Problem for Ordinary Differential Equations). *Let $T \in \mathbb{R}_{>0}$, $u_0 \in \mathbb{R}^n$ and $\Omega \subset \mathbb{R}^n$. Furthermore, let $f : [0, T] \times \Omega \rightarrow \mathbb{R}^n$ be continuous and Lipschitz continuous in the second variable with Lipschitz constant $L > 0$. Then we seek a function $u \in C^1([0, T]; \Omega)$ so that*

$$\begin{aligned} \frac{du}{dt}(t) &= f(t, u(t)) \quad \text{for all } t \in [0, T], \\ u(0) &= u_0. \end{aligned}$$

In the above problem, the smoothness assumptions on f and u are minimal. When introducing and using numerical methods of higher order, we, of course, have to require f and u to be sufficiently smooth. For example Runge-Kutta methods of order p are usually shown to have this convergence order only if f has continuous partial derivatives of at least order p .

To not overcomplicate the demonstration, we restrict ourselves to autonomous ODEs for now. This means that we can write $f(t, u(t)) = f(u(t))$ for all $t \in [0, T]$.

The first way to introduce Rosenbrock-type methods is to linearize the ODE from problem 4.2.1 and then apply Runge-Kutta methods in a certain way. Our illustration here is based on

section 4.5 in the book [28]. If u_m for some $m \in \{0, \dots, M-1\}$ is already computed, we write

$$\begin{aligned} \frac{du}{dt}(t) &= f(u(t)) \\ &= \underbrace{(f(u_m) + J_m(u(t) - u_m))}_{\text{linearization, integrate implicitly}} + \underbrace{(f(u(t)) - f(u_m) - J_m(u(t) - u_m))}_{\text{error, integrate explicitly}} \end{aligned}$$

with the Jacobian $J_m := J_f(u_m) := \left(\frac{\partial f_i}{\partial x_j}(u_m) \right)_{i,j=1}^n$ — so here we already need f to be differentiable in the spatial variables. Then u_{m+1} is computed by integrating the linear term implicitly and the nonlinear error term explicitly with suitable Runge-Kutta methods. Since both explicit and implicit methods integrate constant terms exactly, we can omit the term $f(u_m) - J_m u_m$, which leaves

$$\frac{du}{dt}(t) = \underbrace{J_m u(t)}_{\text{linear}} + \underbrace{(f(u(t)) - J_m u(t))}_{\text{nonlinear}}$$

to be integrated.

Now we show the second way to introduce Rosenbrock-type methods. We will go into more detail here and actually describe the methods completely, including all coefficients. This derivation of the method is largely based on the well-known book [66] on this subject by Strehmel, Weiner and Podhaisky.

Let u and f be “sufficiently” smooth and assume that u_m for some $m \in \{0, \dots, M-1\}$ is already computed. Let $(a_{ij})_{i,j=1}^s$ be the coefficient matrix of an s -stage Diagonally Implicit Runge-Kutta (DIRK) method — in particular we have $a_{ii} \neq 0$ for all $i \in \{1, \dots, s\}$. In the DIRK methods, for each stage $i \in \{1, \dots, s\}$ one has to solve the following — possibly nonlinear — system of equations:

$$k_{mi} = f \left(u_m + \tau_m \sum_{j=1}^i a_{ij} k_{mj} \right).$$

This is usually done with iterative solvers like Newton’s method. In the Rosenbrock-type methods the standard newton’s method is applied/simplified in the following way

- (i) The starting point for the iteration is written as

$$k_{mi}^{(0)} := -\frac{1}{a_{ii}} \sum_{j=1}^{i-1} \gamma_{ij} k_{mj}$$

with another coefficient matrix $(\gamma_{ij})_{i,j=1}^s$.

- (ii) The exact Jacobian $J_f(k_{mi}^{(0)})$ is approximated by a matrix T_{mi} .
- (iii) Only $k_{mi}^{(1)}$ is computed, i.e., only **one** step of the iteration is performed.

This leads to the following **linear** system of equations to be solved for each stage:

$$(I_s - \tau_m a_{ii} T_{mi}) \left(k_{mi}^{(1)} - k_{mi}^{(0)} \right) = -k_{mi}^{(0)} + f \left(u_m + \tau_m \sum_{j=1}^{i-1} a_{ij} k_{mj} + \tau_m a_{ii} k_{mi}^{(0)} \right).$$

With the additional coefficients $(b_i)_{i=1}^s$ and the notations

$$k_{mi} = k_{mi}^{(1)}, \quad \alpha_{ij} = a_{ij} - \gamma_{ij}, \quad \gamma_{ii} = a_{ii}$$

for all $i, j \in \{1, \dots, s\}$. We thus arrive at

Numerical Method 4.2.1 (Rosenbrock-Type Method for Autonomous ODEs).

$$\begin{aligned} (I_s - \tau_m \gamma_{ii} T_{mi}) k_{mi} &= f \left(u_m + \tau_m \sum_{j=1}^{i-1} \alpha_{ij} k_{mj} \right) + \tau_m T_{mi} \sum_{j=1}^{i-1} \gamma_{ij} k_{mj}, \\ &\quad i \in \{1, \dots, s\}, \quad m \in \{0, \dots, M-1\}, \\ u_{m+1} &= u_m + \tau_m \sum_{i=1}^s b_i k_{mi}, \quad m \in \{0, \dots, M-1\} \end{aligned}$$

where $T_{mi} \in \mathbb{R}^{n \times n}$ for $m \in \{0, \dots, M-1\}, i \in \{1, \dots, s\}$, are arbitrary matrices and $(\alpha_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}$, $(\gamma_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}$, $(b_i)_{i=1}^s \in \mathbb{R}^s$ are the coefficients of the method. The right-hand side f and the starting value u_0 are given by the problem 4.2.1.

Because a major motivation for the usage of Rosenbrock-type methods — specifically in the field of parabolic PDEs — is to reduce the amount of linear systems to examine, a first simplification of the above method is to set

$$\gamma_{ii} = \gamma, \quad T_{mi} = T_m \quad \text{for all } i \in \{1, \dots, s\}.$$

This keeps the system matrix constant at each time step.

Using these simplifications and all of the above notations, we will now formulate a simplified Rosenbrock-type method for non-autonomous ODEs, i.e., the right-hand side $f = f(t, u(t))$ also depending explicitly on t .

Numerical Method 4.2.2 (Simplified Rosenbrock-Type Method for Non-Autonomous ODEs).

$$\begin{aligned} (I_s - \tau_m \gamma T_m) k_{mi} &= f \left(t_m + c_i \tau_m, u_m + \tau_m \sum_{j=1}^{i-1} \alpha_{ij} k_{mj} \right) + \tau_m T_m \sum_{j=1}^{i-1} \gamma_{ij} k_{mj} + \tau_m d_i g_m, \\ &\quad i \in \{1, \dots, s\}, \quad m \in \{0, \dots, M-1\}, \\ u_{m+1} &= u_m + \tau_m \sum_{i=1}^s b_i k_{mi}, \quad m \in \{0, \dots, M-1\}, \end{aligned}$$

where $T_m \in \mathbb{R}^{n \times n}$ and $g_m \in \mathbb{R}^n$ for $m \in \{0, \dots, M-1\}$ are arbitrary matrices and vectors, respectively, and $(c_i)_{i=1}^s \in \mathbb{R}^s$, $(d_i)_{i=1}^s \in \mathbb{R}^s$, $(\alpha_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}$, $(\gamma_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}$, $\gamma \in \mathbb{R}_{>0}$, $(b_i)_{i=1}^s \in \mathbb{R}^s$ are the coefficients of the method.

The right-hand side f and the starting value u_0 are given by the problem 4.2.1.

The vectors g_m enter the method as an approximation $g_m \approx \frac{\partial f}{\partial t}(t_m, u_m)$.

In the literature, $\gamma > 0$ is not always required. We include that condition in the definition here because it helps stability and in many cases ensures the invertability of the matrix $I_s - \tau_m \gamma T_m$ — for

example if $T_m = J_f(t_m, u_m) := \left(\frac{\partial f_i}{\partial x_j}(t_m, u_m) \right)_{i,j=1}^n$ and $\frac{du}{dt}(t) = f(t, u(t))$ is a dissipative system.

There are two types of this method that mainly interest us. They are called ROW methods and W-methods, respectively. Initial publications of ROW methods include [25], [72] and [34]. W-methods were first published by Steihaug and Wolfbrandt in the paper [62] and can in some sense be seen as a generalization of ROW methods.

Definition 4.2.1 (ROW Methods). *A method of the type 4.2.2 is called Rosenbrock-Wanner method (in short: ROW method) if*

$$T_m = J_f(t_m, u_m) \quad \text{and} \quad g_m = \frac{\partial f}{\partial t}(t_m, u_m) \quad \text{exactly for all } m \in \{0, \dots, M-1\},$$

as well as

$$c_i = \sum_{j=1}^{i-1} \alpha_{ij} \quad \text{and} \quad d_i = \gamma + \sum_{j=1}^{i-1} \gamma_{ij} \quad \text{for all } i \in \{1, \dots, s\}.$$

ROW methods have the advantage that very stable methods with high order can be constructed, while at the same time the number of order conditions to fulfill can be kept relatively low. The main disadvantage is that the Jacobians have to be computed at each time step.

Definition 4.2.2 (W-Methods). *A method of the type 4.2.2 is called W-Method if*

$$c_i = \sum_{j=1}^{i-1} \alpha_{ij} \quad \text{and} \quad d_i = 0 \quad \text{for all } i \in \{1, \dots, s\}.$$

W-Methods allow the T_m to be arbitrary and thus need to fulfill more order conditions than ROW methods to achieve higher order. They are also not A-stable *in general* (for all possible choices of the T_m) — this can be seen by setting $T_m \equiv 0$ for all $m \in \{0, \dots, M-1\}$, which creates a standard explicit Runge-Kutta method.

The big advantage, however, of W-methods — which makes them so attractive to use — is that the computational cost of calculating Jacobians can be greatly reduced. As we mentioned above, W-methods can be seen as a generalization of ROW methods (excluding the specific value of the d_i). However, we often speak of comparisons and contrasts between ROW methods and W-methods — when doing so, we naturally have W-methods with *inexact* Jacobians in mind.

The specific choice of the T_m in W-methods — albeit not necessarily influencing the order of convergence for small enough time steps — does have a significant impact on computational cost, stability and also accuracy. In practice, the T_m are thus not picked completely arbitrary. A common choice is

$$T_m = J_f(t_m, u_m) + \mathcal{O}(\tau_m),$$

which can be achieved, for example, by keeping the matrix fixed for a number of time steps.

If $f(t, u(t)) = Au(t) + N(t, u(t))$ with A linear and N nonlinear but not causing “too much stiffness” in some sense, a natural choice is to just set

$$T_m = A.$$

Similar choices will be explored when applying W-Methods to semilinear parabolic equations — see section 4.5.1.

4.3 Order, Stability and Dissipation of ODE solvers

The quality of a numerical method can be judged in regard to many different properties — such as order of convergence, stability and others. Its usefulness might also depend on the type of problem that the method is applied to. We are specifically interested in the application of numerical methods to so-called 'stiff' ODEs — this is because, roughly speaking, parabolic PDEs can in some ways be thought of as infinitely stiff ODEs. Hence, many concepts that are used in the treatment of stiff ODEs will be very useful for the semi-discretization in time of parabolic PDEs as well.

Besides usually requiring stable methods, stiff ODEs are also prone to order reduction (a phenomenon that we will investigate in section 4.4) and often make it more difficult to find sharp error estimates. Since stiffness leads to these somewhat unrelated complications and to the author's knowledge, the notion of stiffness has no canonical mathematical definition, we will not try to give a firm definition either.

Rather, we will investigate these complications — that are usually thought of as being related with stiffness — throughout this chapter. We will also introduce two ways to at least gauge how stiff an ODE might be. The first is the stiffness ratio and will be defined just below, the second involves the so-called logarithmic matrix norm and will be talked about briefly in the following section 4.4.

We begin, however, by quickly recalling some standard definitions and results from the field of ODEs and examine how they apply to ROW methods and W-methods.

Definition 4.3.1 (Convergence Order). *We say that a numerical method for the solution of ODEs has convergence order $p \in \mathbb{N}$ if it admits to the following condition when applied to any initial value problem of the type 4.2.1 that has an at least p -times continuously differentiable right-hand side f and the exact solution u .*

For all problems of the above described type, there must exist constants $C > 0$ and $\tau_\infty > 0$ so that for any $M \in \mathbb{N}$ and discrete points in time $0 = t_0 < t_1 < t_2 < \dots < t_{M-1} < t_M \leq T$ with corresponding time step sizes $\tau_\infty \geq \tau_m := t_{m+1} - t_m$, $m \in \{0, \dots, M-1\}$, the numerical method produces approximations $\mathbb{R}^n \ni u_m \approx u(t_m)$, $m \in \{0, \dots, M\}$, which fulfill

$$\max_{0 \leq m \leq M} \|u(t_m) - u_m\| \leq C \left(\max_{0 \leq m \leq M-1} \tau_m \right)^p$$

where $\|\cdot\|$ is some norm on \mathbb{R}^n .

The constants $C > 0$ and $\tau_\infty > 0$ may **not** depend on M or the specific placement of the t_m , $m \in \{0, \dots, M\}$.

There is a lot of literature on the construction of ROW methods and W-methods with arbitrary convergence order. See for example the books [66] or [26] for a detailed derivation of the order conditions that these methods need to fulfill to achieve a certain convergence order. Usually

these conditions are developed through careful comparison of the Taylor expansion of the exact solution u with a Taylor expansion of the numerical solution.

Note that in this presentation we omit the intermediate step of *consistency* order, which together with Lipschitz continuity of f — which we assumed — usually leads to the corresponding convergence order.

Now we list conditions that ensure convergence up to order 3 for ROW methods and W-methods:

Theorem 4.3.1 (Classical Order Conditions for ROW methods and W-Methods). *For ROW methods and W-methods with coefficients as introduced in method 4.2.2 define the following notation for all $i, j \in \{1, \dots, s\}$:*

$$\beta_{ij} := \begin{cases} \alpha_{ij} + \gamma_{ij} & \text{if } i > j \\ \gamma & \text{if } i = j \\ 0 & \text{if } i < j \end{cases}$$

and $\beta_i := \sum_{j=1}^{i-1} \beta_{ij}.$

Remember that for all $i \in \{1, \dots, s\}$ we always have $c_i = \sum_{j=1}^{i-1} \alpha_{ij}$ for ROW methods and W-methods.

The order conditions up to order 3 for ROW methods/W-methods are:

$$\begin{aligned} (O1) \quad & \sum_{i=1}^s b_i = 1, \\ (O2) \quad & \sum_{i=1}^s b_i \beta_i = \frac{1}{2} - \gamma, \\ (O3a) \quad & \sum_{i=1}^s b_i c_i^2 = \frac{1}{3}, \\ (O3b) \quad & \sum_{i,j=1}^s b_i \beta_{ij} \beta_j = \frac{1}{6} - \gamma + \gamma^2, \\ (OW2) \quad & \sum_{i=1}^s b_i c_i = \frac{1}{2}, \\ (OW3a) \quad & \sum_{i,j=1}^s b_i \alpha_{ij} c_j = \frac{1}{6}, \\ (OW3b) \quad & \sum_{i,j=1}^s b_i \alpha_{ij} \beta_j = \frac{1}{6} - \frac{\gamma}{2}, \\ (OW3c) \quad & \sum_{i,j=1}^s b_i \beta_{ij} c_j = \frac{1}{6} - \frac{\gamma}{2}. \end{aligned}$$

- (i) If a ROW method fulfills the first 1/2/4 conditions, it has convergence order 1/2/3, respectively.
- (ii) If a W-method fulfills condition (O1), it has order 1. If it fulfills conditions (O1),(O2),(OW2), it has order 2. If it fulfills all eight of the above listed conditions, it has order 3.
- (iii) If in a W-method $T_m = J_f(t_m, u_m) + \mathcal{O}(\tau_m)$ is used for all $m \in \{0, \dots, M-1\}$, then the conditions (OW3a),(OW3b),(OW3c) are not needed to achieve order 3.

Proofs can be found in the above mentioned books [66] and [26].

Note that the d_i in method 4.2.2 do not enter into the theorems requirements. Therefore, if the coefficients of a Rosenbrock-type method fulfill all requirements to be a *W-method* of order p , then by the above theorem and the definitions 4.2.1 and 4.2.2, the method's coefficients (with possibly adjusted d_i) already fulfill the requirements to be used as a *ROW method* of order p . In that sense, W-methods can also be seen as ROW methods with additional order conditions. In return, they allow the use of arbitrary matrices instead of exact Jacobians.

Another important property of numerical methods is stability. There we are more interested in the behavior of the numerical solution when $t \rightarrow \infty$ and the time step size is bounded from below by a possibly large constant — as opposed to convergence, where time step size approaches zero. Independent of accuracy, a numerical method with medium to large time steps should ideally be applicable on longer time intervals without developing unphysical oscillations or blowing up.

Oftentimes, stability of a method for the solution of ODEs is associated with its applicability to stiff ODEs. For example in [7] one can find the following frequently used concept to describe stiffness:

Definition 4.3.2 (Stiffness Ratio). *Let u be the exact solution of problem 4.2.1 and let the right-hand side f be differentiable in the spatial variables. For each $t \in [0, T]$, let $J(t) := J_f(t, u(t))$. The stiffness ratio of problem 4.2.1 is then defined as*

$$\kappa(t) := \frac{\max_{\lambda \in \sigma_-(t)} |Re \lambda|}{\min_{\lambda \in \sigma_-(t)} |Re \lambda|},$$

where $\sigma_-(t) := \{\lambda \in \mathbb{C} : \lambda \text{ is eigenvalue of } J(t) \text{ and } Re \lambda < 0\}$.

Notice that the stiffness ratio only incorporates eigenvalues with negative real part. This is because eigenvalues with positive real part lead to exponentially increasing solution components that demand small time steps for accuracy reasons in any case.

If the stiffness ratio is large, the problem is said to be stiff, though this might not sufficiently describe the phenomenon of stiffness. There are many criteria by different authors to measure the stiffness of an ODE — we will mention another one of those in section 4.4.

The problem property of a high stiffness ratio leads to increased requirements on numerical solvers. Problems with high stiffness ratio can have smooth and decaying solutions on most of the interval $[0, T]$, so that it would be expected that larger time steps should be usable for most

of the solution process. Some methods, however, require for such problems small time steps on the whole interval to prevent unwanted oscillations or even blow ups of the numerical solution. Those methods lack stability. Roughly speaking, the eigenvalues of the Jacobian with negative real part and high magnitude force those methods to use small time steps, even though the corresponding solution component might already be small.

Since stiffness is such a complex phenomenon, it is only understandable that there are many different kinds of stability: a method can be (strongly) A -stable, (strongly) $A(\alpha)$ -stable, L -stable, B -stable and more.

We want to mention that, of course, all of the above defined methods can also be applied to complex-valued ODEs — i.e., problems, where u and f map into \mathbb{C} . In that case, the matrices T_{mi} , $i \in \{1, \dots, s\}$, $m \in \{0, \dots, M-1\}$ and the vectors g_m , $m \in \{0, \dots, M-1\}$, may be complex-valued as well. Now we briefly present the most commonly known concept of stability — first introduced by Dahlquist in his famous paper [13]. It is sometimes referred to as “linear stability” and relies on the following complex-valued test problem:

Problem 4.3.1 (Dahlquist Test Problem). *Let $\lambda \in \mathbb{C}$. Then we seek a function $u \in C^1(\mathbb{R}_{\geq 0}, \mathbb{C})$ so that*

$$\begin{aligned} \frac{du}{dt}(t) &= \lambda u(t) \quad \text{for all } t \in \mathbb{R}_{\geq 0}, \\ u(0) &= 1. \end{aligned}$$

It is easy to see that the exact solution of that test problem is $u(t) = e^{\lambda t}$ for all $t \in \mathbb{R}_{\geq 0}$.

The idea now is to examine how well a numerical solution $\{u_0 = 1, u_1 \approx u(t_1), \dots, u_M \approx u(T)\}$ with bounded step length

$$\tau_{\text{small}} \leq \max_{0 \leq m \leq M-1} \tau_m \leq \tau_{\text{large}}, \quad \tau_{\text{small}}, \tau_{\text{large}} \in \mathbb{R}_{>0},$$

but T (and thus also M) growing large, mimics the behavior of the before mentioned exact solution. In particular, we have

$$\operatorname{Re} \lambda < 0 \quad \Rightarrow \quad \left(\frac{\partial}{\partial t} |u(t)| < 0 \right) \wedge \left(\lim_{t \rightarrow \infty} |u(t)| = 0 \right)$$

and thus expect

$$\operatorname{Re} \lambda < 0 \quad \Rightarrow \quad (\forall m \in \{0, \dots, M-1\} : |u_{m+1}| < |u_m|) \wedge (|u_M| \rightarrow 0 \text{ as } M \text{ and } T \text{ grow}).$$

The following definition allows us to examine these properties for many numerical methods.

Definition 4.3.3 (Stability Function). *We say that $g : \mathbb{C} \rightarrow \mathbb{C} \cup \{\infty\}$ is the stability function of a given numerical method for the solution of ODEs if for any $M \in \mathbb{N}$, $T \in \mathbb{R}_{>0}$, $0 = t_0 < t_1 < t_2 < \dots < t_{M-1} < t_M \leq T$ and $\tau_m := t_{m+1} - t_m$, $m \in \{0, \dots, M-1\}$ the method, when applied to problem 4.3.1, produces a numerical solution $\{u_m \in \mathbb{R}^n : m \in \{0, \dots, M\}$ that fulfills*

$$u_{m+1} = g(\lambda \tau_m) u_m \quad \text{for all } m \in \{0, \dots, M-1\}.$$

This lets us define the most commonly known concepts of stability.

Definition 4.3.4 (Stability). *Let $g : \mathbb{C} \rightarrow \mathbb{C} \cup \{\infty\}$ be the stability function of a given numerical method for the solution of ODEs. The method is then called*

- (i) A-stable if $|g(z)| \leq 1$ for all $z \in \mathbb{C}$ with $\operatorname{Re} z \leq 0$,
- (ii) strongly A-stable if it is A-stable and $\lim_{\operatorname{Re} z \rightarrow -\infty} |g(z)| < 1$,
- (iii) L-stable if it is A-stable and $\lim_{\operatorname{Re} z \rightarrow -\infty} g(z) = 0$,
- (iv) $A(\alpha)$ -stable for some $\alpha \in (0, \frac{\pi}{2})$ if $|g(z)| \leq 1$ for all $z \in \mathbb{C}$ with $|\arg(z) - \pi| \leq \alpha$,
- (v) strongly $A(\alpha)$ -stable for $\alpha \in (0, \frac{\pi}{2})$ if it is $A(\alpha)$ -stable and $\lim_{\operatorname{Re} z \rightarrow -\infty} |g(z)| < 1$,
- (vi) A_0 -stable if $|g(z)| \leq 1$ for all $z \in \mathbb{R}_{\leq 0}$.

Of course the question remains, what these concepts of stability tell us about the application of numerical methods to larger systems of ODEs — or even PDEs — when they are based on the rather simple test problem 4.3.1. A first answer to that lies in the fact that a system $\frac{du}{dt}(t) = Au(t)$ with a *normal* matrix $A \in \mathbb{C}^{n \times n}$ can essentially be reduced to many instances of problem 4.3.1 with different values for λ . Now if the system $\frac{du}{dt}(t) = Au(t)$ has a high stiffness ratio (see definition 4.3.2) and larger time steps are used, then a strongly A-stable method will produce a smoothly decaying numerical solution for all components that correspond to eigenvalues of A with negative real part. Unstable methods, however, might develop unphysical oscillations or even explode. Thus, our previous definitions of stability are sufficient for this type of linear system as well.

Another property of numerical methods that is related to the stability function, is dissipation. Simply put, that property describes how well a numerical method preserves *real* oscillations of the exact solution instead of dampening them out. If the method has low dissipation, then *physical* and other wanted oscillations are more likely to be preserved. A way to quantify this is to see what the method does, when in the Dahlquist problem 4.3.1, the real part of λ goes to zero. For the exact solutions $\{\mathbb{R}_{\geq 0} \ni t \mapsto e^{\lambda t} : \lambda \in \mathbb{C}\}$ of problem 4.3.1, we have

$$\forall t \in [0, T] \forall b \in \mathbb{R} : \lim_{a \rightarrow 0} |e^{(a+ib)t}| = 1.$$

We want a numerical method to roughly mimic this behavior.

Definition 4.3.5 (Dissipation). *Let $g : \mathbb{C} \rightarrow \mathbb{C} \cup \{\infty\}$ be the stability function of a given numerical method for the solution of ODEs. The method then has low dissipation if for all/some $b \in \mathbb{R}$ there are constants $C_b \geq 1$ so that*

$$\frac{|b|}{C_b} \leq \lim_{a \rightarrow 0} |g(a+ib)| \leq C_b |b|.$$

We say that the closer the constants C_b are to 1 and the more b we can make that claim for, the lower the dissipation of the method is.

A first value to check is $|g(i)|$ — if that is close to 1, then the dissipation of the method is often called decent already. For more details on this subject see section 3.5.1 of [54].

We want to mention that the properties of (strong) A-stability and low dissipation can sometimes be hard to fulfill simultaneously, as, heuristically, A-stability calls for dampening of *unphysical* oscillations, while low dissipation means that *physical* oscillations are preserved.

We will now look at the stability function of ROW methods and W-methods specifically. It is a bit technical, but not particularly difficult to show that the following holds — for a proof see section 8.7.3 of [66].

Proposition 4.3.1. *The stability function of a ROW method with $\gamma > 0$ has for all $z \in \mathbb{C} \setminus \{\frac{1}{\gamma}\}$ the form*

$$g_{ROW}(z) = 1 + zb^T(I_s - z\mathcal{B})^{-1}\mathbf{1}_s, \quad (4.1)$$

where $\mathbf{1}_s = (1, \dots, 1) \in \mathbb{R}^s$ and \mathcal{B} , b originate from the methods coefficients via

$$\mathcal{B}_{ij} := \begin{cases} \alpha_{ij} + \gamma_{ij} & \text{if } i > j \\ \gamma & \text{if } i = j \\ 0 & \text{if } i < j \end{cases}$$

for $i, j \in \{1, \dots, s\}$, as well as $b = (b_i)_{i=1}^s$.

Since \mathcal{B} is a lower triangular matrix with $\mathcal{B}_{ii} = \gamma$ for all $i \in \{1, \dots, s\}$, $I_s - z\mathcal{B}$ is invertible if and only if $z \neq \frac{1}{\gamma}$.

Naturally, the stability function of a W-method using arbitrary approximations to the exact Jacobians will contain expressions involving arbitrary scalars $\tilde{\lambda}$ alongside the given λ from problem 4.3.1. In the literature, stability of a W-method is thus commonly understood as the stability that the method would have if one applied the method to problem 4.3.1 with the exact choice $T_m = \lambda$, $m \in \{0, \dots, M-1\}$.

When using that definition of stability of W-methods, then, of course, the stability functions of W-methods have the exact same form as the stability functions of ROW methods. In the future, we will always have this definition in mind when speaking of A-stable, L-stable, etc. W-methods. There is a lot of literature on the construction of A- and even L-stable ROW methods and W-methods of convergence order up to four — see [64], [26] or [66] for more information.

We want to briefly examine how the stability function of a W-method changes if not the exact λ was used for the $\{T_m : m \in \{0, \dots, M-1\}\}$ but some specific perturbation.

Remark 4.3.1. *For all $m \in \{0, \dots, M-1\}$ let $\delta_m \in \mathbb{C} \setminus \{0\}$ and assume that a W-method with $\gamma > 0$ is applied to problem 4.3.1 with $T_m = \delta_m \lambda$ for all $m \in \{0, \dots, M-1\}$. Then we obtain*

$$u_{m+1} = g_m(\lambda \tau_m) u_m \quad \text{for all } m \in \{0, \dots, M-1\} \text{ with } \lambda \tau_m \neq \delta_m^{-1} \gamma^{-1}$$

and the series of functions

$$\mathbb{C} \setminus \{\delta_m^{-1} \gamma^{-1}\} \ni z \mapsto g_m(z) := 1 + zb^T(I_s - z\mathcal{B}^{(m)})^{-1}\mathbf{1}_s, \quad m \in \{0, \dots, M-1\},$$

where

$$\mathcal{B}_{ij}^{(m)} := \begin{cases} \alpha_{ij} + \delta_m \gamma_{ij} & \text{if } i > j \\ \delta_m \gamma & \text{if } i = j \\ 0 & \text{if } i < j \end{cases}$$

for $i, j \in \{1, \dots, s\}$, as well as $b = (b_i)_{i=1}^s$.

Proof. The application of the W-method to problem 4.3.1 leads to the following systems of equations:

$$\begin{aligned} (I_s - \tau_m \gamma \delta_m \lambda) k_{mi} &= \lambda u_m + \lambda \tau_m \sum_{j=1}^{i-1} \alpha_{ij} k_j + \delta_m \lambda \tau_m \sum_{j=1}^{i-1} \gamma_{ij} k_j \\ &\quad i \in \{1, \dots, s\}, m \in \{0, \dots, M-1\}, \\ \Leftrightarrow (I_s - \tau_m \tilde{\gamma}_m \lambda) k_{mi} &= \lambda u_m + \lambda \tau_m \sum_{j=1}^{i-1} \alpha_{ij} k_j + \delta_m \lambda \tau_m \sum_{j=1}^{i-1} \tilde{\gamma}_{mij} k_j \\ &\quad i \in \{1, \dots, s\}, m \in \{0, \dots, M-1\}, \end{aligned} \quad (4.2)$$

with $\tilde{\gamma}_m := \delta_m \gamma$ and $\tilde{\gamma}_{mij} = \delta_m \gamma_{ij}$ for all $m \in \{0, \dots, M-1\}$, $i, j \in \{1, \dots, s\}$. The new solution is then computed as usual via

$$u_{m+1} = u_m + \tau_m \sum_{i=1}^s b_i k_{mi}, \quad m = 0, \dots, M-1.$$

The equation (4.2) can be interpreted as coming from a method that is similar to a ROW method in that it uses exactly $T_m = \lambda$ for each $m \in \{0, \dots, M-1\}$ but differs from ROW methods because the coefficients γ_m and γ_{mij} change at each time step — which is, of course, not the case for any Rosenbrock-type method. Nonetheless, using the similarity to ROW methods and proposition 4.3.1 the claim from the remark is easily seen. \square

Now, if requirements from definition 4.3.4 could be fulfilled by each individual g_m — at least if the $|\delta_m|$, $m \in \{0, \dots, M-1\}$, are within certain bounds — one could obtain a slightly more accurate description of the true stability of a W-method with arbitrary approximations to the Jacobian. This seems hard to achieve — and most likely is for most methods.

Regarding the ROS2-method, which is an L-stable method introduced in section 4.6, however, we can make the following interesting observations:

For that method, the functions g_m , $m \in \{0, \dots, M-1\}$, from remark 4.3.1 have the form

$$\mathbb{C} \setminus \{\delta_m^{-1} \gamma^{-1}\} \ni z \mapsto g_m(z) = \frac{1 + (1 - 2\delta_m \gamma)z + (0.5 - (1 + \delta_m)\delta_m \gamma + \delta_m^2 \gamma^2)z^2}{(1 - \delta_m \gamma z)^2},$$

with $\gamma = 1 - \frac{1}{\sqrt{2}}$. One can show that if

$$\delta_m \in (0.8990, 1.1548) \quad (4.3)$$

for some $m \in \{0, \dots, M-1\}$, then we have

$$|g_m(z)| < 1 \text{ for all } z \in \mathbb{C} \text{ with } \operatorname{Re} z < 0 \quad (4.4)$$

and

$$\exists \xi \in (0, 1) \exists \epsilon > 0 : |g_m(z)| < \xi \text{ for all } z \in \mathbb{C} \text{ with } \operatorname{Re} z < -\epsilon. \quad (4.5)$$

Thus, if one uses a W-method variant of ROS2 for which all δ_m , $m \in \{0, \dots, M-1\}$, from remark 4.3.1 fulfill condition (4.3), then that W-method variant can more accurately be thought of as being strongly A-stable.

Furthermore, one can also define ROS2 slightly differently by choosing $\gamma = 1 + \frac{1}{\sqrt{2}}$. This also creates an L-stable method with the same convergence order but leads to a somewhat larger error constant (see section 3.1 of [70]). The advantage with that choice of γ is, however, that the δ_m , $m \in \{0, \dots, M-1\}$, may then be from the even larger interval $(0.23682, \infty)$ in order for the g_m to fulfill (4.4) and (4.5).

4.4 Order Reduction and More Accurate Concepts of Convergence for Stiff ODEs

It is well-known that in many numerical experiments with methods for the solution of ordinary differential equations the (classical) convergence order of the method given in definition 4.3.1 is not achieved (see for example section 8.4.1 of [66]). This is another phenomenon attributed to stiffness. Instead, the (numerical) order of convergence drops down to lower integer values — or even fractional order for ODEs that originate from the semi-discretization in space of parabolic PDEs (for a good explanation of the fractional order phenomena see [44]).

To understand why order reduction might happen, let us recall again problem 4.2.1. Many numerical methods for its solution are constructed in a way so that two error bounds can be shown to hold independently of each other:

$$\begin{aligned} \max_{0 \leq m \leq M} \|u(t_m) - u_m\| &\leq C_1(L)\tau^p \\ \text{and} \quad \max_{0 \leq m \leq M} \|u(t_m) - u_m\| &\leq C_2(\nu)\tau^q, \end{aligned}$$

with $p, q \in \mathbb{N}$, $p > q$, $\tau := \max_{0 \leq m \leq M-1} \tau_m$ and constants $C_1(L), C_2(\nu) > 0$.

Here we assume that $C_1(L)$ depends on the given Lipschitz constant L of the right-hand side f of problem 4.2.1, while $C_2(\nu)$ does **not** depend on L but (among other dependencies) on a so-called *one-sided Lipschitz constant* $\nu \in \mathbb{R}$ of f , a term that will be defined below. For now it suffices to know that in many stiff problems, L is very large, while ν can be of moderate magnitude or even negative. For a lot of numerical methods the corresponding constants $C_1(L), C_2(\nu)$ can be shown to be of very different size for those stiff problems.

Lets assume for a moment that $C_2(\nu)$ is of moderate size, but $C_1(L) \gg 1$ is very large. Then, the first (classical) of the above error bounds will not be sharp for many step sizes and will not give an accurate representation of the convergence behavior for those step sizes. We have

$$C_1(L)\tau^p \leq C_2(\nu)\tau^q \quad \Leftrightarrow \quad \tau \leq \left(\frac{C_2(\nu)}{C_1(L)} \right)^{\frac{1}{p-q}}.$$

Since for many numerical methods and stiff problems L — and thus also $C_1(L)$ — can indeed be very large independently of $C_2(\nu)$, this leads to an impractically small step length being required to achieve the (classical) order p . Rather, the observed order for all relevant values of τ then

is q at most. This also happens for *stable* methods. Thus, order reduction is another problem originating from stiff problems that needs its own treatment when trying to construct numerical methods of higher order.

In general, one is interested in deriving error estimates which correspond to the behavior of the *exact solution* u of problem 4.2.1 on the whole interval $[0, T]$. For example, if u is smooth, one would ideally expect the estimated error to be rather small as well. Classical error estimates usually depend on the *right-hand side* f and global bounds for the norms of its derivatives — these bounds might not reflect the behavior of the solution on the whole of $[0, T]$, though, if the problem is stiff.

Before we look at an example, we introduce some terms and results that are commonly used in regard to order reduction.

Definition 4.4.1 (One-Sided Lipschitz Condition). *Let (\cdot, \cdot) be a dot-product on \mathbb{R}^n and $\|\cdot\|$ the associated norm. Let $T \in \mathbb{R}_{>0}$. A function $f : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ then admits to a one-sided Lipschitz condition if there exists a constant $\nu \in \mathbb{R}$ — called one-sided Lipschitz constant of f — so that*

$$(f(t, y) - f(t, v), y - v) \leq \nu \|y - v\|^2$$

for all $t \in [0, T]$ and $y, v \in \mathbb{R}^n$.

We promptly make the following

Remark 4.4.1.

- (i) *By the Cauchy-Schwartz inequality, a function $f : [0, T] \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ that is Lipschitz continuous in the second variable with Lipschitz constant L admits to a one-sided Lipschitz condition with one-sided Lipschitz constant $\nu = L$. There may, however, be other one-sided Lipschitz constants smaller than $\nu = L$, of course.*
- (ii) *The one-sided Lipschitz constant can be negative. If 4.2.1 has a right-hand side which has a non-positive one-sided Lipschitz constant, the problem is said to be dissipative and has a 'stable' solution in the sense that slightly differing initial values lead to only small differences in the corresponding solutions. For more details see section 7.2 of [66].*

As already mentioned above, the one-sided Lipschitz constant of a given function f can be positive and small — or even negative — while the classical Lipschitz constant of f might be very large. The following definition and the subsequent results allow us to better understand this in the case of linear systems.

Definition 4.4.2 (Logarithmic Matrix Norm). *Let $\|\cdot\|$ be a norm on \mathbb{R}^n . Let $A \in \mathbb{R}^{n \times n}$ and denote by $\|A\|$ the corresponding induced matrix norm. Then the limes*

$$\mu[A] := \lim_{\delta \rightarrow 0^+} \frac{\|I_n + \delta A\| - 1}{\delta}$$

is called the logarithmic norm of A induced by $\|\cdot\|$.

For the logarithmic norms induced by $\|\cdot\|_p$, $p \in \mathbb{N} \cup \{\infty\}$, we write $\mu_p[\cdot]$.

Next, we demonstrate the connection to the one-sided Lipschitz constant and summarize other important properties of the logarithmic matrix norm.

Remark 4.4.2.

- (i) Let $\|\cdot\|$ be a norm on \mathbb{R}^n and $A \in \mathbb{R}^{n \times n}$. Then the induced logarithmic norm $\mu[A]$ exists and we have

$$\max_{j \in \{1, \dots, r\}} \operatorname{Re} \lambda_j \leq \mu[A] \leq \|A\|,$$

where $\{\lambda_1, \dots, \lambda_r\}$, $r \in \{1, \dots, n\}$, are the eigenvalues of A .

- (ii) Let (\cdot, \cdot) be a dot-product on \mathbb{R}^n and $\|\cdot\|$ the associated norm. If $f(t, u)$ from problem 4.2.1 is continuously differentiable in the spatial variables, then we have for the logarithmic norm induced by $\|\cdot\|$ and for any $v \in \mathbb{R}^n$

$$\mu[J_f(t, y)] \leq \nu \quad \text{for all } t \in [0, T], y \in \mathbb{R}^n,$$

if and only if

$$(f(t, y) - f(t, v), y - v) \leq \nu \|y - v\|^2 \quad \text{for all } t \in [0, T], y, v \in \mathbb{R}^n.$$

- (iii) For any $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n \times n}$ we have

$$\begin{aligned} \mu_1[A] &= \max_{j \in \{1, \dots, n\}} \left(a_{jj} + \sum_{i \in \{1, \dots, n\} \setminus \{j\}} |a_{ij}| \right) \\ \mu_\infty[A] &= \max_{i \in \{1, \dots, n\}} \left(a_{ii} + \sum_{j \in \{1, \dots, n\} \setminus \{i\}} |a_{ij}| \right) \\ \mu_2[A] &= \lambda_{\max} \left(\frac{1}{2} (A + A^T) \right) \end{aligned}$$

where $\lambda_{\max}(\frac{1}{2}(A + A^T))$ is the maximum over all eigenvalues of $\frac{1}{2}(A + A^T)$.

Proofs can be found in section 7.2 of [66].

Based on the logarithmic norm, there is another characterization of stiffness, which is also from [66] — specifically from section 7.4 of that book. One can say that problem 4.2.1 with continuously differentiable right-hand side and at least two solution components is stiff if for some norm $\|\cdot\|$ and the induced logarithmic norm $\mu[\cdot]$ we have

$$\mu[J_f] < -\|J_f\| \quad \text{and} \quad T \|J_f\| > 1.$$

According to some authors however, this still does not encompass all situations where stiffness occurs (see [5] for a detailed explanation). Namely, stiffness can also occur, when J_f is time-dependent and not close to a normal matrix in some sense. In that case it is possible that $1 < \mu[J_f] \approx \|J_f\|$ even though the solution of the corresponding system is smooth after a short initial phase. This can lead to error bounds, which depend only on a one-sided Lipschitz constant

and derivatives of the exact solution that are still not sharp on the whole interval.

To illustrate the concepts of the logarithmic norm and the one-sided Lipschitz condition, let us now look at a simple one-dimensional heat equation.

Problem 4.4.1 (Initial Boundary Value Problem for the One Dimensional Heat Equation). *Let $T \in \mathbb{R}_{>0}$ and $v_0 \in C^2([0, 1], \mathbb{R})$. Then we seek a function $v : [0, T] \times [0, 1] \rightarrow \mathbb{R}$ that is continuously differentiable in the first variable on $[0, T]$, twice continuously differentiable in the second variable on $[0, 1]$ and fulfills*

$$\begin{aligned} \frac{\partial v}{\partial t}(t, x) &= \frac{\partial^2 v}{\partial x^2}(t, x) && \text{for all } (t, x) \in (0, T] \times (0, 1), \\ v(t, 0) &= v(t, 1) = 0 && \text{for all } t \in [0, T], \\ v(0, x) &= v_0(x) && \text{for all } x \in [0, 1]. \end{aligned}$$

We first discretize this problem in space using a central difference quotient, i.e., by choosing $N \in \mathbb{N}$ with $N > 2$ and setting $h := \frac{1}{N}$ we approximate

$$\frac{\partial^2 v}{\partial x^2}(t, x) \approx \frac{1}{h^2} (v(t, x+h) - 2v(t, x) + v(t, x-h))$$

for all $(t, x) \in (0, T] \times (0, 1)$. Alongside setting $u_0(t) = v(t, 0) = 0$ and $u_N(t) = v(t, 1) = 0$ exactly for all $t \in [0, 1]$, the solution to problem 4.4.1 can then be approximated as $v(t, jh) \approx u_j(t)$ for all $t \in [0, T]$ and $j \in \{1, \dots, N-1\}$, where $u = (u_j)_{j=1}^{N-1}$ is a solution to the following system of ordinary differential equations:

Problem 4.4.2 (Initial Value Problem for the Spatially Discretized 1D Heat Equation). *We seek a function $u \in C^1([0, T]; \mathbb{C}^{N-1})$ so that*

$$\begin{aligned} \frac{du}{dt}(t) &= Au(t) && \text{for all } t \in [0, T], \\ u(0) &= (v_0(jh))_{j=1}^{N-1}, \end{aligned}$$

with

$$A = \frac{1}{h^2} \begin{pmatrix} -2 & 1 & & & 0 \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ 0 & & & 1 & -2 \end{pmatrix} \in \mathbb{R}^{N-1 \times N-1}$$

and $v_0 \in C^2([0, 1], \mathbb{R})$ given by problem 4.4.1.

It is not hard to show that for small h , the eigenvalue of A with the largest absolute value is approximately $-\frac{4}{h^2}$ and the one with the smallest absolute value is approximately $-\pi^2$. That shows us that for small h

- (a) the stiffness ratio (see definition 4.3.2) of A is approximately $\frac{4}{\pi^2 h^2}$,
- (b) the spectral norm $\|A\|_2$ is approximately $\frac{4}{h^2}$ and thus any Lipschitz constant of the right-hand side of problem 4.4.2 is at least of that size as well,

- (c) the logarithmic norm $\mu_2[A]$ is approximately $-\pi^2$ so the right-hand side of problem 4.4.2 has a one-sided Lipschitz constant of around $-\pi^2$ (see remark 4.4.2).

Overall, we have an ordinary differential equation whose stiffness ratio and Lipschitz constants become very large — nearing infinity — when the spatial discretization becomes finer. This significantly restricts the pool of suitable numerical methods for the solution of problem 4.4.2. Methods that fulfill no stability requirements (see definition 4.3.4) will usually have an impractical time step size restriction of the form $\tau \lesssim h^2$. On the other hand, when using methods of higher order, the large Lipschitz constant in conjunction with the small one-sided Lipschitz constant can lead to the aforementioned order reduction.

Roughly speaking, $-A$ is a “discrete version” of the Dirichlet Laplacian, which is an unbounded operator with unbounded but countable spectrum that lies entirely on the positive real axis and has a minimum greater than zero. The eigenvalues of A resemble this. The better $-A$ approximates the Dirichlet Laplacian (i.e., the finer the spatial discretization gets), the larger $\|A\|_2$ — and thus also any Lipschitz constant — becomes, while $\mu_2[A]$ and appropriate one-sided Lipschitz constants converge to $-\pi^2$.

Naturally, we will have to deal with similar problems whenever solving parabolic PDEs — as those generally involve unbounded operators with spectral properties comparable to that of the Dirichlet Laplacian.

We want to briefly address the question how a substantial difference $\nu \ll L$ between one-sided and classical Lipschitz constant can lead to different error estimates — and thus order reduction — as claimed at the beginning of this section. We will not go into the technicalities here.

The answer lies in the different ways in which Taylor expansion is used to derive error estimates. One way requires the partial derivatives of the right-hand side f , which in turn introduce the classical Lipschitz constant L into the error estimates. Another way applies Taylor expansion only to the exact solution and thus requires only derivatives of u — in conjunction with the one-sided Lipschitz condition 4.4.1 and the one-sided Lipschitz constant ν , this can be sufficient to show convergence without having the error constants depend on L . Now if ν is positive and small or even negative, the error estimate that only depends on ν and derivatives of u can be sharp, while the error estimate that depends on $L \gg \nu$ might not be — details on how exactly these ‘stiff error estimates’ are derived can be found for example in chapter IV section 15 of [26].

Such error estimates give rise to the concepts of B -convergence (first introduced by Frank, Schneid and Ueberhuber in [17]) and B_{PR} -convergence. Those concepts are attempts to give rigorous definitions of convergence that are better suited for stiff equations. We want to emphasize again, though, that error estimates which only depend on ν and derivatives of u are still not necessarily sharp in all cases (see the discussion below remark 4.4.2).

Definition 4.4.3 (B-Convergence). *We say that a method is B-convergent of order $q \in \mathbb{N}$ if it has (classical) convergence order q in the sense of definition 4.3.1 **and** the constants $C > 0$ and $\tau_\infty > 0$ from that definition do **not** directly depend on partial derivatives of right-hand sides f of problems of the type 4.2.1 — in particular not on (classical) Lipschitz constants of those right-hand sides. They **may**, however, depend on one-sided Lipschitz constants ν of right-hand sides and also on derivatives of the (unique) exact solutions u .*

Quite a lot of research has been done on determining under which conditions a method can be B-convergent of some order and which methods fulfill these conditions. Runge-Kutta methods with higher B-convergence order can be constructed — see again chapter IV section 15 of [26] for more details.

Unfortunately, it can be shown that ROW methods and W-methods can never be B-convergent *at all* — once more, chapter IV section 15 of [26] contains details on that. One can, however, show 'stiff convergence' of ROW methods and W-methods for a smaller class of problems. B-convergence was defined for *all* problems with right-hand sides that are Lipschitz continuous and fulfill a one-sided Lipschitz condition. Two subsets of these problems are of main interest to us. The first one is

Problem 4.4.3 (Semilinear ODE). *We seek a solution u to a modified problem 4.2.1, where the right-hand side f is defined to be of the following form:*

There exist $A \in \mathbb{R}^{n \times n}$, $g : [0, T] \times \Omega \rightarrow \mathbb{R}^n$ and constants $\xi \in \mathbb{R}$, $L_g \in \mathbb{R}_{>0}$ so that for all $v, y \in \mathbb{R}^n$ and $t \in [0, T]$ it holds that

$$f(t, v) = Av + g(t, v)$$

and

$$\begin{aligned} \mu[A] &\leq \xi, \\ \|g(t, v) - g(t, y)\| &\leq L_g \|v - y\|, \end{aligned} \tag{4.6}$$

with the logarithmic norm $\mu[A]$ being induced by a norm $\|\cdot\|$ that is associated with a dot-product (\cdot, \cdot) on \mathbb{R}^n .

The second one is

Problem 4.4.4 (Prothero-Robinson). *We seek a solution u to a modified problem 4.2.1 where the right-hand side f is defined to be of the following form:*

There exist $\lambda \in \mathbb{R}_{<0}$ and a function $g \in C^1([0, T], \mathbb{R})$ so that for all $v \in \mathbb{R}^n$ and $t \in [0, T]$ it holds that

$$f(t, v) = \lambda(v - g(t)) + g'(t).$$

Obviously, problem 4.4.4 is a semilinear problem of the type 4.4.3.

We first observe the following:

Remark 4.4.3.

(i) *Let f be a right-hand side of problem 4.4.3. We then get for all $v, y \in \mathbb{R}^n$ and $t \in [0, T]$*

$$\begin{aligned} \|f(t, v) - f(t, y)\| &\leq (\|A\| + L_g) \|v - y\| \quad \text{and} \\ (f(t, v) - f(t, y), v - y) &\leq (A(v - y), v - y) + L_g \|v - y\|^2 \leq (\mu[A] + L_g) \|v - y\|^2, \end{aligned}$$

where (\cdot, \cdot) , $\|\cdot\|$, $\mu[A]$ are as in problem 4.4.3 and $\|A\|$ is the matrix norm induced by $\|\cdot\|$. The last inequality follows with (ii) of remark 4.4.2.

(ii) Let f be a right-hand side of problem 4.4.4. We then get for any dot-product (\cdot, \cdot) on \mathbb{R}^n and its associated norm $\|\cdot\|$ that for all $v, y \in \mathbb{R}^n$, $t \in [0, T]$ the following holds

$$\begin{aligned} \|f(t, v) - f(t, y)\| &\leq |\lambda| \|v - y\| \quad \text{and} \\ (f(t, v) - f(t, y), v - y) &= (\lambda(v - y), v - y) = \lambda \|v - y\|^2. \end{aligned}$$

Thus, we see that the right-hand sides in the above defined problems have continuity properties such that those problems really are general ODE problems of the type 4.2.1. Furthermore, this shows that by common characterizations of stiffness, the semilinear problem 4.4.3 is stiff if L_g is of moderate size and $\mu[A] \ll 0$.

Because the Prothero-Robinson problem 4.4.4 just contains a scalar equation, none of our characterizations of stiffness can be applied to it. This makes sense in a way because for $g(t) \equiv 0$ and $v(0) = 1$, we just get Dahlquist's problem 4.3.1, which usually is not considered stiff for any value of λ . If one sets $v(0) = g(0)$, however, then the solution of the Prothero-Robinson problem 4.4.4 is exactly $v(t) = g(t)$ independent of λ . If in addition one chooses for g a smoothly decaying function, such as $g(t) = e^{-t}$, then unstable, explicit methods will not produce useful solutions for large time steps and $\lambda \ll 0$, even though the exact solution is the same for all values of λ .

Furthermore, many numerical methods do not achieve their classical order of convergence when applied to the Prothero-Robinson problem 4.4.4 with $\lambda \ll 0$. Thus, for some choices of $v(0)$, g and λ , problem 4.4.4 should certainly be considered stiff. In the author's opinion, the following — mathematically not precise — description gives a good understanding of stiffness that also fits the Prothero-Robinson problem 4.4.4. The description is inspired by the book [66] (see section 7.4 of that book) as well:

An ordinary differential equation is stiff if explicit solvers require small step sizes for stability reasons — even though the solution does not change significantly — while implicit solvers allow for larger step sizes. In other words, the equation is stiff if the step length used in explicit solvers is influenced by stability rather than accuracy.

The idea is that the semilinear problem 4.4.3 covers many stiff problems that arise in applications, while the Prothero-Robinson problem 4.4.4 represents the core of what makes many stiff problems so hard to handle (the original paper of Prothero and Robinson [48] has a detailed motivation for choosing problem 4.4.4 as a test problem). Hence, we will now examine in more detail, under which conditions ROW methods and W-methods can be B-convergent for these smaller problem classes. Furthermore, it turns out that fulfilling some of those conditions will also be beneficial when semi-discretizing in time or fully-discretizing parabolic PDEs. We begin by presenting the utilized definitions.

Definition 4.4.4 (B-Convergence for Semilinear ODEs). *We say that a method is B-convergent of order $q \in \mathbb{N}$ for problems of the type 4.4.3 if for some $q \in \mathbb{N}$ it fulfills the modified version of definition 4.4.3 where the problem type 4.2.1 is replaced by the problem type 4.4.3.*

In a similar way we define B-convergence for the Prothero-Robinson problem 4.4.4. However, the simpler problem structure allows a more specific definition that gives a better picture of what happens when simultaneously $\lambda\tau \rightarrow -\infty$ and $\tau \rightarrow 0$, i.e., when the problem becomes 'infinitely stiff'.

Definition 4.4.5 (B-Convergence for the Prothero-Robinson Problem). *We say that a method is B-convergent of the order $q \in \mathbb{N}$ for problems of the type 4.4.4 (in short: B_{PR} -convergent of order q) if for any problem of the type 4.4.4 with exact solution u , there exist constants $C_1, C_2 \geq 0$, $\tau_\infty > 0$ independent of the given problem parameter λ , so that the application of the method to that problem produces approximations $\mathbb{R}^n \ni u_m \approx u(t_m)$, $m \in \{0, \dots, M\}$, which fulfill the error estimate*

$$\max_{0 \leq m \leq M} \|u(t_m) - u_m\| \leq C_1 \tau^p + C_2 \frac{\tau^{p+1}}{|z|},$$

where $\tau_\infty \geq \tau := \max_{0 \leq m \leq M-1} \tau_m$ and $z := \lambda \min_{0 \leq m \leq M-1} \tau_m$ with the time step sizes $\tau_0, \dots, \tau_{M-1}$.

Notice that the constants C_1, C_2, τ_∞ do not depend on the specific value of λ (though they might depend on the specific function g given by problem 4.4.4). This makes sense because we only allowed negative real values for λ , so requiring the error constants to not depend on λ corresponds to having them be independent of the stiffness of the problem.

This definition — with the term $\frac{1}{|z|}$ in the error estimate — is taken from [51]. It can be shown for some methods that the error can indeed be bounded by $\frac{\tau^{p+1}}{|z|}$. For those methods, the above definition emphasizes good convergence behavior if simultaneously $|z| \rightarrow \infty$ and $\tau \rightarrow 0$.

The examination of B-convergence for semilinear problems of the type 4.4.3 is somewhat complicated. It turns out that under some restrictive conditions on the coefficients, Rosenbrock-type methods with higher B-convergence order for problems of the type 4.4.3 can be constructed — details are in [65] and again [66] (in that latter book, see specifically remark 11.6.3.). Another reason why we will focus on B_{PR} -convergence instead, is that some interesting applications from the field of parabolic PDEs do not really match the semilinear ODE problem 4.4.3 — one could say that a “PDE-version” of the Lipschitz condition (4.6) would often be too strict.

Before we introduce what will be the main set of conditions that we want our methods to fulfill in addition to classical order conditions, we want to briefly talk about how B-convergence on the problems 4.4.3 and 4.4.4 is occasionally examined for W-methods.

To be able to give rigorous mathematical proofs, it is sometimes assumed that the matrix T of the W-method is not completely arbitrary, but set to

$T = A$ for the semilinear problem 4.4.3 or $T = \lambda$ for the Prothero-Robinson problem, 4.4.4

respectively. The idea is that in real applications T will be chosen so that it is in some sense ‘not too far away’ from either A or — for more general problems — the real Jacobian. If that is the case, the hope is that B-convergence conditions developed for the simplified choices $T = A$ for problem 4.4.3 or $T = \lambda$ for problem 4.4.4 will still help to prevent order reduction for inexact Jacobians and problems that are not even semilinear problems of the type 4.4.3. Notice that the choice $T = A$ for problem 4.4.3 means already that not the exact Jacobian, but a very specific replacement is used. In section 4.5.5. of the book [64], the choice

$T = A + P_m$ in problem 4.4.3 with a matrix P_m depending on the current time step index m is examined. In that book, Strehmel and Weiner present additional requirements on P_m and the W-method that allow higher order B-convergence for problems of the type 4.4.3.

Now we list an established set of conditions that ensure B_{PR} -convergence of higher order for ROW methods. They stem from a series of papers by Rang — with [50] and [51] being two recent ones. From sections 2 and 4.1 of [51], a proof of the following result can be obtained.

Theorem 4.4.1 (Conditions for B_{PR} -Convergence of ROW Methods). *For a ROW method of the form 4.2.2 we use the additional notation from theorem 4.3.1 and to further shorten the presentation we define*

$$b := (b_1, \dots, b_s)^T, \quad d := (d_1, \dots, d_s), \quad \mathcal{B} := (\beta_{ij})_{i,j=1}^s$$

and $c^k := (c_1^k, \dots, c_s^k)^T$ for all $k \in \mathbb{N} \cup \{0\}$.

Also, remember that for ROW methods we require $d_i = \gamma + \sum_{j=1}^{s-1} \gamma_{ij}$ for all $i \in \{1, \dots, s\}$.

If a ROW method fulfills all (classical) order conditions up to order $p \geq 2$ (see theorem 4.3.1) and in addition it fulfills the conditions

$$\forall k \in \{2, \dots, p\} : \quad b^T \mathcal{B}^{-1} c^k = 1, \quad (4.7)$$

$$\forall k \in \{3, 4, \dots\} \forall l \in \{\max\{1, k-p\}, \dots, k-2\} :$$

$$b^T \mathcal{B}^{-(l+1)} \frac{1}{k-l} c^{k-l} = b^T \mathcal{B}^{-l} (c^{k-l-1} + d \delta_{k-l-1,1}), \quad (4.8)$$

where $\delta_{k-l-1,1}$ is the usual Kronecker, then the method is

- (i) B_{PR} -convergent of order $p-1$ if it is A -stable,
- (ii) B_{PR} -convergent of order p if it is strongly A -stable and equidistant time steps are used,
- (iii) B_{PR} -convergent of order p if it is L -stable.

Condition (4.8) specifically addresses error terms of the form $\frac{\tau^{k-l}}{|z|^l}$. If (4.8) is not fulfilled by a method, its convergence for medium stiff problems, i.e., $|\lambda|$ not too large, could be poor. For a detailed examination of the dominance of different error terms in different regimes of λ , see again [51] and earlier papers by that author.

For B_{PR} -convergence of order 2, the conditions in the above result may be replaced by another condition, which was given and used in several papers such as [29], [58] or [45].

Theorem 4.4.2 (Conditions for B_{PR} -Convergence of Order 2 of ROW Methods). *Here we use the notations from theorem 4.4.1.*

If a ROW method fulfills all (classical) order conditions up to order 2 and in addition it fulfills the condition

$$\forall k \in \{2, \dots, s+1\} : \quad b^T \mathcal{B}^{k-1} (c + d) = \frac{1}{2} b^T \mathcal{B}^{k-2} c^2, \quad (4.9)$$

then the method is

- (i) B_{PR} -convergent of order 1 if it is A -stable,

- (ii) B_{PR} -convergent of order 2 if it is strongly A -stable and equidistant time steps are used,
- (iii) B_{PR} -convergent of order 2 if it is L -stable.

A proof of this can be assembled from section 2 of [50] and section 2 of [51].

As it turns out, condition (4.9) can also be used to increase the convergence order in time of the discretization of SPEs via ROW methods.

The conditions (4.7), (4.8) and (4.9) will be the main set of conditions that we want our higher order ROW methods and W-methods to fulfill in addition to the classical order conditions. In the next section we will move from ODEs to SPEs and use ROW methods and W-methods to discretize them in time. For ROW methods, condition (4.9) is used to rigorously prove better convergence. In numerical experiments with Prothero-Robinson problems and parabolic PDEs, the other conditions (4.7) and (4.8) have lead to improved numerical results — see for example the papers [50, 51] by Rang. Hence, we will be looking for ROW methods and W-methods that also fulfill the conditions (4.7), (4.8), even though we do not have a strict mathematical argument regarding the usefulness of those conditions for the discretization of SPEs.

4.5 ROW Methods and W-Methods for SPEs

Remember that in the semilinear parabolic problem 3.7.1 we are looking on the time interval $[0, T]$, $T \in \mathbb{R}_{>0}$, for a function $u \in C([0, T]; H)$ with $u(t) \in V$, $\frac{du}{dt}(t) \in V'$ for all $t \in [0, T]$ and

$$\begin{aligned} \frac{du}{dt}(t) + \mathbb{A}u(t) &= N(u(t)) + G(t) \quad \text{for all } t \in (0, T], \\ u(0) &= u_0 \in H. \end{aligned}$$

Here, (V, H, V') is a triplet of complex Hilbert spaces, $\mathbb{A} : D(\mathbb{A}) \subset H \rightarrow H$ is a sectorial operator which can be extended to a bounded operator from V to V' , $N : V \rightarrow V'$ is the nonlinearity and $G : (0, T] \rightarrow V'$ is the forcing term.

Now we present the discretization in time of this problem via Rosenbrock-type methods. We will simply replace the finite dimensional objects (Matrices, vectors, (nonlinear) functions) in method 4.2.2 with the corresponding — potentially infinite dimensional — Hilbert space objects from problem 3.7.1.

We still work with the assumption formulated at the beginning of this chapter, which is that there are $M + 1 \in \mathbb{N}$ discrete points

$$0 = t_0 < t_1 < t_2 < \dots < t_{M-1} < t_M \leq T,$$

taken from the time interval $[0, T]$, $T \in \mathbb{R}_{>0}$, with the corresponding time step sizes $\tau_m := t_{m+1} - t_m$, $m \in \{0, \dots, M-1\}$. For a given unique solution u to the above problem, we then seek an approximation $\{u_m \in H : m \in \{0, \dots, M\}\}$, i.e., $u_m \approx u(t_m)$ for all $m \in \{0, \dots, M\}$.

For sufficiently differentiable N and G in the above problem, we will look to iteratively solve the linear equations in the following numerical method in order to find such an approximation.

Numerical Method 4.5.1 (ROW method/W-method for Semilinear Parabolic Equations).

$$\begin{aligned}
 (1 + \tau_m \gamma \mathbb{T}_m) k_{mi} &= G(t_m + c_i \tau_m) - \mathbb{A} \left(u_m + \tau_m \sum_{j=1}^{i-1} \alpha_{ij} k_{mj} \right) + N \left(u_m + \tau_m \sum_{j=1}^{i-1} \alpha_{ij} k_{mj} \right) \\
 &\quad - \tau_m \mathbb{T}_m \sum_{j=1}^{i-1} \gamma_{ij} k_{mj} + \tau_m d_i g_m, \quad i \in \{1, \dots, s\}, m \in \{0, \dots, M-1\}, \\
 u_{m+1} &= u_m + \tau_m \sum_{i=1}^s b_i k_{mi}, \quad m \in \{0, \dots, M-1\},
 \end{aligned}$$

where $(c_i)_{i=1}^s \in \mathbb{R}^s$, $(d_i)_{i=1}^s \in \mathbb{R}^s$, $(\alpha_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}$, $(\gamma_{ij})_{i,j=1}^s \in \mathbb{R}^{s \times s}$, $\gamma \in \mathbb{R}_{>0}$, $(b_i)_{i=1}^s \in \mathbb{R}^s$ are the coefficients of the method.

The linear operator \mathbb{A} , the functions N , G and the starting value u_0 are given by the problem 3.7.1.

- (i) For ROW methods we have for all $m \in \{0, \dots, M-1\}$ exactly $g_m = \frac{dG}{dt}(t_m)$ and $\mathbb{T}_m = \mathbb{A} - D^{(1)}N(u_m)$, where $D^{(1)}N$ is the Fréchet derivative of N . Naturally, this means that N and G need to be sufficiently differentiable.

The coefficients of ROW methods fulfill

$$c_i = \sum_{j=1}^{i-1} \alpha_{ij} \quad \text{and} \quad d_i = \sum_{j=1}^i \gamma_{ij} \quad \text{for all } i \in \{1, \dots, s\}.$$

- (ii) For W-methods the $\mathbb{T}_m : V \rightarrow V'$ can be any linear operators and the coefficients fulfill

$$c_i = \sum_{j=1}^{i-1} \alpha_{ij} \quad \text{and} \quad d_i = 0 \quad \text{for all } i \in \{1, \dots, s\}.$$

As already mentioned at the end of section 4.2, the \mathbb{T}_m in the W-methods will usually not be completely arbitrary. A significant difficulty with those methods is finding the right strategy for determining the \mathbb{T}_m . That strategy should lead to accurate (no order reduction, low absolute error), stable (no blow up of longer simulations) and cheap (low computing cost) schemes. Balancing these requirements will be one of our main tasks when using these methods in fully-discrete algorithms in chapter 6.

In the following convergence analysis, the error of the above introduced temporal discretization in comparison with the exact solution is given in various norms of potentially infinite dimensional Hilbert spaces. We will use the $\|\cdot\|_V$, the $\|\cdot\|_H$ and the $\|\mathbb{A}^\alpha \cdot\|_V = \|\cdot\|_{\mathbb{A}^\alpha}$, $\alpha \in (0, 1)$ norms. In numerical experiments with parabolic PDEs, we cannot observe these errors exactly — as, of course, we always have to discretize in space as well. Semi-discrete error bounds, however, let us isolate the temporal error and, roughly speaking, they cover the case that a spatial mesh size nears zero and thus — in a way — the temporal stiffness of the corresponding discrete equations nears infinity.

When testing these semi-discrete error estimates in practice, we will usually keep the spatial

discretization as fine as possible — and even choose solutions which are in the discrete spaces at each point in time — to minimize the spatial error.

Error bounds for the time discretization of problem 3.7.1 are normally generated by developing the exact solution u into a Taylor series, which is then compared to the numerical solution. Consequently, the error bound will depend on the temporal smoothness of the exact solution.

This approach is similar to the one we introduced in the previous section, where B-convergence of ODEs was established as a notion of convergence that depends on the temporal smoothness of the exact solution but not on derivatives and Lipschitz constants of the right-hand side. For problems of the type 3.7.1 this makes even more sense because the (possibly unbounded) operator \mathbb{A} is most likely not Lipschitz continuous as a mapping from H to H or V to V . Therefore we cannot necessarily expect convergence of order p of the time discretization by a ROW method or a W-method if the method has classical order p . Further order conditions, very similar to the ones from the previous section, will be necessary to achieve higher order of convergence for the time discretization of problem 3.7.1.

In the 1990s, Lubich and Ostermann developed in a series of papers a convergence theory suited for the discretization in time of problem 3.7.1 and other types of problems. We do not restate all their results here. For example we largely omit their sharp error estimates that show the fractional order of temporal convergence which can be induced by spatial smoothness and boundary conditions. We instead refer the interested reader to the papers [40], [42] and [41] for more details.

From [41] we get this result for ROW methods:

Theorem 4.5.1. *Let $q \in \{2, 3\}$ and assume that problem 3.7.1 has a unique solution u on the whole of $[0, T]$ with at least the temporal regularity $u \in H^{q+1}(0, T; V)$.*

Now we augment problem 3.7.1 with a number of additional requirements:

- (i) *Let the first and second Fréchet derivatives $D^{(1)}N : V \rightarrow \mathcal{L}(V, V')$ and $D^{(2)}N : V \rightarrow \mathcal{L}^2(V \times V, V')$ of the nonlinearity exist. Furthermore, for all $r \in \mathbb{R}_{>0}$ let there be a constant $C_{D^{(2)}N, r} > 0$ so that for all $v \in V$ with $\|v\|_V \leq r$ and all $w_1, w_2 \in V$ we have*

$$\|(D^{(2)}N(v))(w_1, w_2)\|_{V'} \leq C_{D^{(2)}N, r} \|w_1\|_V \|w_2\|_V.$$

- (ii) *Let there be time-independent constants $d \in \mathbb{R}$, $\epsilon > 0$, $\phi \in (0, \frac{\pi}{2})$ and $C_{\tilde{\mathbb{A}}, inv}, C_{\tilde{\mathbb{A}}}, M_{\tilde{\mathbb{A}}}, L_{\tilde{\mathbb{A}}} > 0$ so that the operators $\{\tilde{\mathbb{A}}(t) := \mathbb{A} - D^{(1)}N(u(t)) + d : t \in [0, T]\}$ (remember that u is the exact solution of problem 3.7.1) have the following properties:*

- (a) *$\tilde{\mathbb{A}}(t)$ is invertible for all $t \in [0, T]$ with*

$$\|\tilde{\mathbb{A}}(t)\|_{\mathcal{L}(V, V')} \leq C_{\tilde{\mathbb{A}}} \quad \text{and} \quad \|\tilde{\mathbb{A}}(t)^{-1}\|_{\mathcal{L}(V', V)} \leq C_{\tilde{\mathbb{A}}, inv}.$$

- (b) *For all $t \in [0, T]$ and all $\lambda \in -S_{\epsilon, \phi}$ the operator $(\lambda + \tilde{\mathbb{A}}(t)) : V \rightarrow V'$ is invertible with*

$$\|(\lambda + \tilde{\mathbb{A}}(t))^{-1}\|_{\mathcal{L}(V, V)} \leq \frac{M_{\tilde{\mathbb{A}}}}{1 + |\lambda|}.$$

(c) The mapping $[0, T] \ni t \mapsto \tilde{\mathbb{A}}(t) \in \mathcal{L}(V, V')$ is differentiable with

$$\|\tilde{\mathbb{A}}(t_1) - \tilde{\mathbb{A}}(t_2)\|_{\mathcal{L}(V, V')} \leq L_{\tilde{\mathbb{A}}} |t_1 - t_2| \quad \text{for all } t_1, t_2 \in [0, T].$$

Then there exist constants $\tau_{ROW}, C_{ROW} \in \mathbb{R}_{>0}$ so that any ROW method that

- is strongly $A(\alpha)$ -stable for some $\alpha > \phi$,
- uses equidistant time steps with some time step size $\tau \leq \tau_{ROW}$ (i.e., for some $M \in \mathbb{N}$ with $M \leq \frac{T}{\tau}$, the points $\{t_m = \frac{mT}{M} : m \in \{0, \dots, M\}\}$ are used),
- is applied to problem 3.7.1, which is augmented with the above described additional requirements,
- has at least (classical) order q and, if $q = 3$, admits to the additional condition (4.9),

yields only uniquely solvable equations and produces an approximate solution $\{u_m \in V : m \in \{0, \dots, M\}\}$ that fulfills the following:

$$\left(\tau \sum_{m=0}^M \|u_m - u(t_m)\|_V^2 \right)^{\frac{1}{2}} + \max_{0 \leq m \leq M} \|u_m - u(t_m)\|_H \leq C_{ROW} \tau^q. \quad (4.10)$$

The constants C_{ROW}, τ_{ROW} depend on the parameters of the ROW method, on all the constants from the requirements (i) and (ii) of this theorem and on the exact solution u .

As the lengthy formulation already indicates, the proof of this theorem is quite technical and intricate. It uses a theory of perturbed Rosenbrock-type methods and employs concepts such as Taylor expansion of the exact solution, Peano kernels, sectorial operators and their resolvent bounds, generating functions, and others. In the source [41], the complete proof uses multiple results from that paper and from the paper [42]. We did attempt to merge all these results into one proof and somewhat simplify it, but ultimately failed with the simplification and thus decided to not present the long proof here. Instead, we focus on the applicability of the above result (and a similar one for W-methods, see theorem 4.5.2 below) to specific parabolic semilinear equations in the following section 4.5.1.

At first glance, the above result contains extensive requirements on the equation that might not be fulfilled very often. We will see, however, that in many cases — such as the classic examples we already introduced — those requirements are not an issue. Before we look at those examples, we turn our attention to W-methods.

Naturally, it is even harder to prove convergence when arbitrary operators \mathbb{T}_m are used instead of exact derivatives. In the same paper [41] from which we got the previous theorem, Lubich and Ostermann proved a result that provides convergence up to order 2 if the operators \mathbb{T}_m fulfill certain requirements. To keep the presentation as clear and concise as possible, we simplify their setup but keep it broad enough to cover the choices for the \mathbb{T}_m that interest us. Below the convergence result, we will look at some of those choices.

Theorem 4.5.2. Assume that problem 3.7.1 has a solution u on the whole of $[0, T]$ with at least the temporal regularity $u \in H^3(0, T; V)$.

Now we augment problem 3.7.1 with the requirements (i) and (ii) from the previous theorem 4.5.1. We further assume that there exists some $\beta \in [0, 1]$ so that the following additional requirements hold:

- (i) For the operators $\{\tilde{\mathbb{A}}(t) : t \in [0, T]\}$ introduced in requirement (ii) of theorem 4.5.1, there is a constant $C_{\tilde{\mathbb{A}}, V} > 0$ so that for all $t_1, t_2 \in [0, T]$ we have $\tilde{\mathbb{A}}(t_1)^{-\beta}(V) = \tilde{\mathbb{A}}(t_2)^{-\beta}(V)$ and $\frac{1}{C} \|\tilde{\mathbb{A}}(t_1)^\beta v\|_V \leq \|\tilde{\mathbb{A}}(t_1)^\beta v\|_V \leq C \|\tilde{\mathbb{A}}(t_1)^\beta v\|_V$ for all $v \in \tilde{\mathbb{A}}(t_1)^{-\beta}(V)$. Furthermore, the exact solution has improved spatial regularity in the sense that $\frac{du}{dt} \in L^2(0, T; (\tilde{\mathbb{A}}(t)^{-\beta}(V), \|\tilde{\mathbb{A}}(t)^\beta \cdot\|_V))$ for some (and thus all) $t \in [0, T]$.
- (ii) There exist constants $\tilde{c} \in \mathbb{R}$, $\tilde{\epsilon} > 0$, $\tilde{\phi} \in (0, \frac{\pi}{2})$ and $C_{\mathbb{T}}, C_{\mathbb{T}, inv}, M_{\mathbb{T}}, C_{\mathbb{A}, \mathbb{T}}, C_{\mathbb{T}, V}, C_{\mathbb{T}, V'} > 0$ so that for all $M \in \mathbb{N}$ and all $m \in \{0, \dots, M-1\}$ there exists a linear operator $\mathbb{T}_{M,m} \in \mathcal{L}(V, V')$ with the subsequent properties:

- (a) The operator $\tilde{\mathbb{T}}_{M,m} := \mathbb{T}_{M,m} + \tilde{c}$ is invertible with

$$\|\tilde{\mathbb{T}}_{M,m}\|_{\mathcal{L}(V, V')} \leq C_{\mathbb{T}} \quad \text{and} \quad \|\tilde{\mathbb{T}}_{M,m}^{-1}\|_{\mathcal{L}(V', V)} \leq C_{\mathbb{T}, inv}.$$

Further, for all $\lambda \in -S_{\tilde{\epsilon}, \tilde{\phi}}$ the operator $(\lambda + \tilde{\mathbb{T}}_{M,m}) : V \rightarrow V'$ is also invertible and we have

$$\|(\lambda + \tilde{\mathbb{T}}_{M,m})^{-1}\|_{\mathcal{L}(V, V)} \leq \frac{M_{\tilde{\mathbb{T}}}}{1 + |\lambda|}.$$

- (b) For all $v \in V$ we have

$$\|\tilde{\mathbb{T}}_{M,m}^\beta \tilde{\mathbb{A}}(t_m)^{-\beta} v\|_V \leq C_{\mathbb{A}, \mathbb{T}} \|v\|_V, \quad (4.11)$$

$$\|\mathbb{E}_{M,m} \tilde{\mathbb{T}}_{M,m}^{-\beta} v\|_V \leq C_{\mathbb{T}, V} \|v\|_V, \quad (4.12)$$

$$\|\mathbb{E}_{M,m} \tilde{\mathbb{T}}_{M,m}^{-\beta} v\|_{V'} \leq C_{\mathbb{T}, V'} \|v\|_{V'}, \quad (4.13)$$

where

$$t_m := \frac{mT}{M} \quad \text{and} \quad \mathbb{E}_{M,m} := \tilde{\mathbb{T}}_{M,m} - \tilde{\mathbb{A}}(t_m).$$

Then there exist constants $\tau_W, C_W \in \mathbb{R}_{>0}$ so that any W -method that

- is strongly $A(\alpha)$ -stable for some $\alpha > \phi$,
- uses equidistant time steps with some time step size $\tau \leq \tau_W$ (i.e., for some $M \in \mathbb{N}$ with $M \leq \frac{T}{\tau}$, the points $\{t_m = \frac{mT}{M} : m \in \{0, \dots, M\}\}$ are used),
- is applied to problem 3.7.1 which is augmented with the above described additional requirements,
- uses approximate operators $\{\mathbb{T}_{M,m} \in \mathcal{L}(V, V') : m \in \{0, \dots, M-1\}\}$ as described in requirement (ii) of this theorem,
- has at least (classical) order 2,

yields only uniquely solvable equations and produces an approximate solution $\{u_m \in V : m \in \{0, \dots, M\}\}$ that fulfills the following:

$$\left(\tau \sum_{m=0}^M \|u_m - u(t_m)\|_V^2 \right)^{\frac{1}{2}} + \max_{0 \leq m \leq M} \|u_m - u(t_m)\|_H \leq C_W \tau^2. \quad (4.14)$$

The constants C_W, τ_W depend on the parameters of the W-method, on all the constants from the requirements (i) and (ii) of this and the previous theorem and on the exact solution u .

If one were to prove this result, the major difference compared to ROW methods would be that the approximative operators lead to more complicated expressions in error recursion defects of the perturbed W-methods. The resolvent bound in (ii) (a) and the conditions (4.11), (4.12), (4.13) can be used to bound those defects. Once again, though, we do not present Lubich and Ostermann's long proof here, but instead focus on the applicability of the result to specific equations.

Notice that the operators $\{\tilde{\mathbb{T}}_{M,m} : M \in \mathbb{N}, m \in \{0, \dots, M-1\}\}$ in requirement (ii) of the above theorem are by part (a) of that requirement sectorial operators on V' (with domain V) with well-defined fractional powers. Hence, the expressions in (4.11), (4.12) and (4.13) make sense.

Another significant thing to see is that, of course, the t_m , introduced in part (b) of requirement (ii), really do also depend on M — as do the u_m in the numerical solution. For the sake of brevity, we suppress that dependence. We do not do so for the $\mathbb{T}_{M,m}$ and $\mathbb{E}_{M,m}$ to emphasize that all the constants in requirement (ii) should not depend on M . Otherwise the constant C_W in the bound at the end of the theorem could depend on M and thus on the step size, possibly leading to no convergence at all.

Requirement (ii) of the above result describes which kind of operators we want to allow in our W-methods instead of the exact Fréchet derivative of $\mathbb{A} + N$ at the numerical solution (those exact Fréchet derivatives would be used in ROW methods, of course). By (ii) (a), these operators need to preserve most properties of \mathbb{A} . Indeed, as we will see, $\mathbb{T}_{M,m} := \mathbb{A}$ for all $M \in \mathbb{N}$ and all $m \in \{0, \dots, M-1\}$ is often a valid and sensible choice.

When the involved operators are differential operators — and that is always the case in our examples — then inequality (4.11) can roughly be understood as $\mathbb{T}_{M,m}$ **not** having higher order derivatives than $\mathbb{A}(t_m)$ for all $m \in \{0, \dots, M-1\}$. Similarly, the inequalities (4.12) and (4.13) describe that the “operator-errors” $E_{M,m}$, $m \in \{0, \dots, M-1\}$, have only derivatives up to an order that depends on β . In our applications we will often work with $\beta = \frac{1}{2}$ and elliptic operators, which means that the $E_{M,m}$ only contain derivatives of order 1 or less. A larger value for β obviously means that (4.12) and (4.13) are easier to fulfill, but it also raises the regularity requirements for the exact solution via requirement (i) of the above result.

Lastly, we want to mention that the above result is formulated for methods of classical order $p \geq 2$ and as such might not be sharp for all problems if methods with higher classical order are used. Ideally, one would extend the above convergence result to W-methods of higher order — eventually arriving at a sharp fractional order of convergence which depends on spatial smoothness and boundary conditions.

However, we do not know of any result in the style of theorem 4.5.2 that is formulated for W-methods of order 3 or greater which are applied to SPEs. As mentioned above, there are results like that for ROW methods and many Runge-Kutta methods — see the papers [41] and [42]. The paper [61] contains some further convergence bounds for W-methods applied to semilinear equations, but those bounds are formulated in different norms and in a slightly different setting than ours.

4.5.1 Application of ROW Methods and W-Methods to Some Specific SPEs

In this section, we want to examine whether and how the above introduced convergence results can be applied to the convection-diffusion equation, the incompressible Navier-Stokes equations and some variants of convection-diffusion equations with zero-order nonlinearity. Again, in all of the following arguments, $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, is a bounded domain with “sufficiently smooth” boundary and all function spaces are *complex*-valued.

We also want to mention once more that in the following applications we always assume the existence of a unique solution to the described problems that has at least H^3 -regularity in time and exists on the whole time interval. We will then sometimes state higher regularity requirements if necessary.

We start with ROW methods and theorem 4.5.1 as its requirements are a bit easier to examine.

ROW Methods in the Linear Case

In problem 3.7.1 let $\mathbb{A} \in \{\mathbb{D}, \mathbb{S}\}$ with \mathbb{D} and \mathbb{S} defined as in sections 3.3 and 3.4, respectively, and let $N \equiv 0$. Then obviously **requirement (i)** in the above theorem 4.5.1 is fulfilled. Since there is no time-dependence in the operators, **requirement (ii)** follows directly from the sections 3.3 and 3.4 in conjunction with section 3.6.

ROW Methods for the Incompressible Navier-Stokes Problem

In problem 3.7.1 let $\mathbb{A} := \nu \mathbb{S}$ with some $\nu > 0$ and \mathbb{S} defined as in section 3.4, and let $N(v) := (- (v \cdot \nabla) v, \cdot)_{L^2} \in [V_\sigma(\Omega)]'$ for all $v \in V = V_\sigma(\Omega)$ as in section 3.8. Then it is easily seen that for all $v, w_1, w_2 \in V$ we have

$$(D^{(2)}N(v))(w_1, w_2) = - (w_1 \cdot \nabla) w_2 - (w_2 \cdot \nabla) w_1$$

independent of v . By (i) of theorem 3.8.2, **requirement (i)** of theorem 4.5.1 is then fulfilled.

For the exact solution u of problem 3.7.1 we make the additional regularity assumption of $u \in C([0, T]; H_\alpha(\mathbb{S}))$ for some $\alpha \in (\frac{3}{4}, 1]$. From theorem 3.8.4 we then get **requirement (ii) (a) and (b)**.

Since $u \in H^{q+1}(0, T; V)$ for some $q \in \{2, 3\}$ is already a requirement of theorem 4.5.1, the Bochner space embedding from theorem 3.1.2 gives us in particular $u \in C^1([0, T]; V)$. Using (c) of (i) of theorem 3.8.2, we get **requirement (ii) (c)** by observing

$$\|D^{(1)}N(u(t_1)) - D^{(1)}N(u(t_2))\|_{\mathcal{L}(V, V')} \leq 2\|u(t_1) - u(t_2)\|_V \leq 2\left\|\frac{du}{dt}\right\|_{C([0, T]; V)} |t_1 - t_2|$$

for all $t_1, t_2 \in [0, T]$.

ROW Methods for Some Reaction-Diffusion Problems

In problem 3.7.1 let $\mathbb{A} := \mathbb{D}$ with \mathbb{D} defined as in section 3.3 and for some $r \in \mathbb{N}$ and some $(a_k)_{k=0}^r \in \mathbb{R}^r$ let $N(v) := \sum_{k=0}^r a_k v^k$ for all $v \in V = H_0^1(\Omega)$. Depending on the size of r , we can show that the requirements of theorem 4.5.1 are fulfilled in this situation.

If, for example, we have $N(w) = w^r$ for $r \in \{2, 3\}$ and all $w \in V$, then we can show for all $v, w_1, w_2 \in V$ that

$$\begin{aligned} \|N(v)\|_{L^2} &= \|v^r\|_{L^2} \leq \tilde{C}_0 \|v\|_{L^6}^r \leq C_0 \|\nabla v\|_{L^2}^r, \\ \|(D^{(1)}N(v))(w_1)\|_{L^2} &= \|rv^{r-1}w_1\|_{L^2} \leq r\tilde{C}_1 \|v\|_{L^6}^{r-1} \|w_1\|_{L^6} \\ &\leq rC_1 \|\nabla v\|_{L^2}^{r-1} \|\nabla w_1\|_{L^2}, \end{aligned} \tag{4.15}$$

$$\begin{aligned} \|(D^{(2)}N(v))(w_1, w_2)\|_{L^2} &= \|r(r-1)v^{r-2}w_1w_2\|_{L^2} \leq r(r-1)\tilde{C}_2 \|v\|_{L^6}^{r-2} \|w_1\|_{L^6} \|w_2\|_{L^6} \\ &\leq r(r-1)C_2 \|\nabla v\|_{L^2}^{r-2} \|\nabla w_1\|_{L^2} \|\nabla w_2\|_{L^2}, \end{aligned} \tag{4.16}$$

where the constants $C_0, C_1, C_2 > 0$ only depend on Ω . First of all this shows that for $r \leq 3$, N really maps into V' and that $D^{(1)}N$ and $D^{(2)}N$ exist everywhere. By inequality (4.16), **requirement (i)** of theorem 4.5.1 is immediately given.

One might think that r could be taken larger than 3 because we only need $N, D^{(1)}N, D^{(2)}N$ to map V into V' , not $L^2(\Omega)$. However, to obtain **requirement (ii) (a) and (b)** by generalizing the convection-diffusion operator to the operators $\{\mathbb{D} - D^{(1)}N(u(t)) : t \in [0, T]\}$ uniformly in time — with $u \in C([0, T]; V)$ being our exact solution — we need

$$\|D^{(1)}N(u(t))w\|_{L^2} \leq C\|\nabla w\|_{L^2} \quad \text{for all } w \in V \tag{4.17}$$

with some time-independent constant C . So inequality (4.15) with its L^2 -bound really is required. To show that bound, we used Hölder's inequality with the Sobolev embedding $H^1(\Omega) \subset L^6(\Omega)$, so there we really do use $r \leq 3$. The reason why we need the bound in the first place is that we work with very similar techniques as those that were used to generalize the Stokes to the time-dependent Oseen operator uniformly in time — with the bound (4.17) essentially just replacing the bound (3.17) from the proof of the first Oseen result 3.8.3 in section 3.8.1. Since the steps are so similar, we omit the details here.

We obtain **requirement (ii) (c)** of theorem 4.5.1 very similarly to the way it was shown in the above subsection on ROW methods for Navier-Stokes — again we note that the regularity $u \in C^1([0, T]; V)$ is gained through the Bochner space embeddings from theorem 3.1.2 and

$u \in H^{q+1}(0, T; V)$ for some $q \in \{2, 3\}$ already being a requirement of theorem 4.5.1.

Now we check the requirements of theorem 4.5.2. The main difficulty for W-methods will be, of course, that we also need to specify with which operators we want to replace the exact Fréchet derivatives that would be used in ROW methods. The inspection of those operator choices will make it significantly more intricate and lengthy to verify the requirements of theorem 4.5.2 compared to our previous examination of the ROW method theorem 4.5.1.

As a somewhat easy but still interesting example, we start with the convection-diffusion equation and choose the Laplacian as an approximation of the full convection-diffusion operator.

W-Methods in the Linear Case

In problem 3.7.1 let $N \equiv 0$ and $\mathbb{A} := \mathbb{D}$, where \mathbb{D} is defined as in section 3.3 with the diffusion coefficient $\nu > 0$, the convection direction $b \in L^\infty(\Omega, \mathbb{R})^d$ and the reaction coefficient $c \in L^\infty(\Omega, \mathbb{R})$. We then know from the above subsection on ROW methods for the linear case that all conditions of theorem 4.5.1 are fulfilled in this situation. In particular, there is a $d \in \mathbb{R}$ so that $\tilde{\mathbb{D}} := \mathbb{D} + d$ has 0 in its resolvent set and can be extended to a bounded, invertible operator in $\mathcal{L}(H_0^1(\Omega), H^{-1}(\Omega))$ with bounded inverse and well-defined fractional powers.

Requirement (i) in theorem 4.5.2 is obviously fulfilled for any $\beta \in [0, 1]$ if the exact solution is sufficiently smooth.

Now set $\mathbb{T}_{M,m} := \mathbb{L}$ for all $M \in \mathbb{M}$ and all $m \in \{0, \dots, M-1\}$. **Requirement (ii) (a)** follows from sections 3.3 and 3.6.

We will show in more detail that **requirement (ii) (b)** is met as well. The proof relies heavily on the H^2 -regularity result for elliptic equations and the characterization of fractional spaces of elliptic operators as Sobolev spaces.

Because we want to allow first order derivatives in the “operator-errors” (see the discussion below theorem 4.5.2) we choose $\beta := \frac{1}{2}$. For all $v \in H_0^1(\Omega)$ we have $\tilde{\mathbb{D}}^{-\frac{1}{2}}v = \tilde{\mathbb{D}}^{-1}\tilde{\mathbb{D}}^{\frac{1}{2}}v \in D(\tilde{\mathbb{D}}) = H_0^1(\Omega) \cap H^2(\Omega)$ and thus get:

$$\begin{aligned} \|\mathbb{L}^{\frac{1}{2}}\tilde{\mathbb{D}}^{-\frac{1}{2}}v\|_{H^1} &\leq C_{\mathbb{D}1}^{(0)}\|\mathbb{L}\tilde{\mathbb{D}}^{-\frac{1}{2}}v\|_{L^2} \leq C_{\mathbb{D}1}^{(1)}\|\tilde{\mathbb{D}}^{-\frac{1}{2}}v\|_{H^2} \\ &\leq C_{\mathbb{D}1}^{(2)}\|\tilde{\mathbb{D}}^{\frac{1}{2}}v\|_{L^2} \leq C_{\mathbb{D}1}^{(3)}\|v\|_{H^1} \end{aligned}$$

where all constants — in particular the last one, $C_{\mathbb{D}1}^{(3)} > 0$ — are independent of v . Thus, the first part of requirement (ii) (b) is shown.

For the second part of requirement (ii) (b) we assume the regularity $b \in W^{1,\infty}(\Omega, \mathbb{R})^d$ for the transport direction and $c \in W^{1,\infty}(\Omega, \mathbb{R})$ for the zero-order coefficient. We want to mention here that for a larger β , a bit less regularity for b and c , such as just H^1 -regularity, might be sufficient.

We obtain the following for all $v \in H_0^1(\Omega)$ and $x := \mathbb{L}^{-\frac{1}{2}}v \in H^2(\Omega)$ by using the product

rule of differentiation:

$$\begin{aligned} \|(b \cdot \nabla)x + cx\|_{H^1} &\leq C_{\mathbb{D}2}^{(0)} (\|b\|_{W^{1,\infty}} \|x\|_{H^1} + \|b\|_{L^\infty} \|x\|_{H^2} + \|c\|_{W^{1,\infty}} \|x\|_{H^1} + \|c\|_{L^\infty} \|x\|_{H^2}) \\ &\leq C_{\mathbb{D}2}^{(1)} \|x\|_{H^2} \leq C_{\mathbb{D}2}^{(2)} \|\mathbb{L}x\|_{L^2} \leq C_{\mathbb{D}2}^{(3)} \|\mathbb{L}^{\frac{1}{2}}x\|_{H^1} = C_{\mathbb{D}2}^{(3)} \|v\|_{H^1}, \end{aligned}$$

where all constants — in particular the last one, $C_{\mathbb{D}2}^{(3)} > 0$ — are independent of v .

For the third part of requirement (ii) (b), we again assume $b \in W^{1,\infty}(\Omega, \mathbb{R})^d$. Using integration by parts for the transport term, we get for all $y, w \in H_0^1(\Omega)$:

$$\begin{aligned} |\langle (b \cdot \nabla)y + cy, w \rangle| &= |((b \cdot \nabla)y + cy, w)_{L^2}| = |-(w \operatorname{div} b + (b \cdot \nabla)w + cw, y)_{L^2}| \\ &\leq (\|b\|_{W^{1,\infty}} \|w\|_{L^2} + \|b\|_{L^\infty} \|\nabla w\|_{L^2} + \|c\|_{L^\infty} \|w\|_{L^2}) \|y\|_{L^2} \\ &\leq C_{\mathbb{D}3}^{(0)} \|w\|_{H^1} \|y\|_{L^2} \end{aligned}$$

with some constant $C_{\mathbb{D}3}^{(0)} > 0$ independent of y, w .

Furthermore, we have for all $y \in H_0^1(\Omega)$:

$$\begin{aligned} \|y\|_{L^2} &= \sup_{\substack{h \in L^2(\Omega) \\ h \neq 0}} \frac{|(y, h)_{L^2}|}{\|h\|_{L^2}} = \sup_{\substack{w \in H_0^1(\Omega) \\ w \neq 0}} \frac{|(y, \mathbb{L}^{\frac{1}{2}}w)_{L^2}|}{\|\mathbb{L}^{\frac{1}{2}}w\|_{L^2}} \leq C_{\mathbb{D}3}^{(1)} \sup_{\substack{w \in H_0^1(\Omega) \\ w \neq 0}} \frac{|(y, \mathbb{L}^{\frac{1}{2}}w)_{L^2}|}{\|w\|_{H^1}} \\ &= C_{\mathbb{D}3}^{(1)} \sup_{\substack{w \in H_0^1(\Omega) \\ w \neq 0}} \frac{|(\mathbb{L}^{\frac{1}{2}}y, w)_{L^2}|}{\|w\|_{H^1}} = C_{\mathbb{D}3}^{(1)} \sup_{\substack{w \in H_0^1(\Omega) \\ w \neq 0}} \frac{|\langle \mathbb{L}^{\frac{1}{2}}y, w \rangle|}{\|w\|_{H^1}} = C_{\mathbb{D}3}^{(1)} \|\mathbb{L}^{\frac{1}{2}}y\|_{H^{-1}} \end{aligned}$$

with some constant $C_{\mathbb{D}3}^{(1)} > 0$ independent of y .

By setting $y := \mathbb{L}^{-\frac{1}{2}}v \in H_0^1(\Omega)$ for all $v \in H_0^1(\Omega)$ and using the two inequalities above, we obtain the third part of requirement (ii) (b).

Next, we will look at the semilinear Navier-Stokes equation. In a W-method with $M \in \mathbb{N}$ steps, we will replace the exact Fréchet derivative (which would be used in ROW methods) at the numerical solution vectors $\{u_m : m \in \{0, \dots, M-1\}\}$ with the Fréchet derivative at some other vectors $\{\hat{u}_{M,m} : m \in \{0, \dots, M-1\}\}$.

In order to show that requirement (ii) of theorem 4.5.2 is fulfilled, we will work with the assumption (4.18), which states that the $\hat{u}_{M,m}$ are bounded in H^2 uniformly in M and m . This setup certainly includes the case of updating the Jacobian only once — at the beginning — if the initial data is regular enough. Here we use the word Jacobian for the Fréchet derivative of $\mathbb{S} + N$. The case of sporadically updating the Jacobian is *most likely* covered by this setup as well, though we do not have a strict proof of that at the moment — see the discussion below for more details.

W-Methods for the Incompressible Navier-Stokes Problem

In problem 3.7.1 let $\mathbb{A} := \nu \mathbb{S}$ with some $\nu > 0$ and \mathbb{S} defined as in section 3.4, and let $N(v) := (-(v \cdot \nabla)v, \cdot)_{L^2} \in [V_\sigma(\Omega)]'$ for all $v \in V_\sigma(\Omega)$ as in section 3.8. Now let u be a

sufficiently smooth exact solution of the semilinear Navier-Stokes problem. We mostly require $u \in C([0, T]; H_\alpha(\mathbb{S}))$ for some $\alpha \in (\frac{3}{4}, 1]$ — for a few of the following arguments we need even more regularity, though.

From the above subsection on ROW methods for Navier-Stokes we get that for some $d \in \mathbb{R}$, the operators $\hat{A}(t) := \nu \mathbb{S} - D^{(1)}N(u(t)) + d[= \mathbb{O}(t) + d]$, $t \in [0, T]$, fulfill requirements (i) and (ii) of theorem 4.5.1 if $u \in C([0, T]; H_\alpha(\mathbb{S})) \cap C^1([0, T]; V_\sigma(\Omega))$ for some $\alpha \in (\frac{3}{4}, 1]$.

Because we again want to allow first order derivatives in the “operator-errors”, we set $\beta := \frac{1}{2}$. By our first theorem 3.8.3 on the Oseen operator, **requirement (i)** of theorem 4.5.2 is then fulfilled if the exact solution has the regularity $u \in H^1(0, T; H^2(\Omega)^d \cap V_\sigma(\Omega))$.

For all $M \in \mathbb{N}$, $m \in \{0, \dots, M-1\}$ and some $\hat{u}_{M,m} \in V_\sigma(\Omega)$ we now set

$$\mathbb{T}_{M,m}v := \nu \mathbb{S}v + P_\sigma((\hat{u}_{M,m} \cdot \nabla)v + (v \cdot \nabla)\hat{u}_{M,m})$$

for all $v \in H^2(\Omega)^d \cap V_\sigma(\Omega)$. In order to be able to show that **requirement (ii)** is fulfilled, we always assume

$$\exists C_{\mathbb{T}, H^2} > 0 \forall M \in \mathbb{N} \forall m \in \{0, \dots, M-1\} : \|\hat{u}_{M,m}\|_{H^2} \leq C_{\mathbb{T}, H^2}. \quad (4.18)$$

The assumption is fulfilled, of course, if $\hat{u}_{M,m}$ is from a fixed finite subset of $D(\mathbb{S}) = H^2(\Omega)^d \cap V_\sigma(\Omega)$ for all $M \in \mathbb{N}$ and all $m \in \{0, \dots, M-1\}$.

The simplest — but also a useful — choice where this is the case is $\hat{u}_{M,m} = 0$ for all $M \in \mathbb{N}$ and all $m \in \{0, \dots, M-1\}$.

Another interesting case is the one, where we want to update the Jacobian every now and then but not necessarily in every step. For our operators, the numerical solution given by method 4.5.1 is by well-known elliptic regularity results always from $D(\mathbb{S})$ if the data is smooth enough. Since we can, of course, only calculate finitely many numerical solution steps, a possible finite set to choose the $\hat{u}_{M,m}$ from is the union of all numerical solution steps for all step sizes. Using this finite set does include the cases that we update the Jacobian not necessarily in every step but only in every fifth or tenth step or when some indicator tells us to.

However, the constant $C_{\mathbb{T}, H^2}$ (and with it the constant C_W from theorem 4.5.2) could then, of course, grow if the number of different step sizes and/or the number of times we update the Jacobian grows. If we want $C_{\mathbb{T}, H^2}$ to stay relatively small, independent of the step size and the times we update the Jacobian, we need some kind of H^2 -stability result for the numerical solution. To the author’s knowledge, no such result is currently known. A similar stability result, however, for the simpler case of linear equations and a constant choice $\mathbb{T}_{M,m} = \mathbb{T}$ for all $m \in \{0, \dots, M-1\}$, can be found in the paper [43] by Ostermann.

Furthermore, numerical results indicate that this choice of $\mathbb{T}_{M,m} = \nu \mathbb{S} - D^{(1)}N(u_{\iota(m)})$, where $u_{\iota(m)}$ is a solution from some time step previous to the m th of M steps, does indeed produce a stable solution if the step size is small enough. Lastly, it seems intuitive that if the case of updating the Jacobian only once (in the beginning) is covered by the theory, that updating the Jacobian more often should be even better, and that is also almost exclusively what the numerical experiments exhibit.

Thus, we will from now on always work under the assumption (4.18) and expect that updating

the Jacobian every now and then is covered by it. Nevertheless, giving a rigorous H^2 -stability proof under suitable requirements would certainly be helpful and could be a goal for the future. Another possible approach would be to look for different convergence proofs for W-methods that use sporadic Jacobian updates. One could try to obtain a convergence proof without needing all the strong requirements of requirement (ii) of theorem 4.5.2.

Because of assumption (4.18), the operators $\{\mathbb{T}_{M,m} : M \in \mathbb{N}, m \in \{0, \dots, M-1\}\}$ are Oseen-type operators which can be extended (using the results from section 3.6) to operators that fulfill **requirement (ii) part (a)** of theorem 4.5.2. This can be proven by replacing the exact Navier-Stokes solution $\{u(t) : t \in [0, T]\}$ in the Oseen operator definition with the $\{\hat{u}_{M,m} : M \in \mathbb{N}, m \in \{0, \dots, M-1\}\}$ and then using the exact same arguments as were used in section 3.8.1 on the Oseen operator to obtain all the “ (M, m) -uniform” results that we need.

In particular, there is a constant $\tilde{c} \in \mathbb{R}$ — independent of M and m — so that the operators $\tilde{\mathbb{T}}_{M,m} := \mathbb{T}_{M,m} + \tilde{c}$ are sectorial operators with 0 in their respective sectors and “ (M, m) -uniform” resolvent bounds on those sectors. This implies (just as in section 3.8.1) that these operators can be “ (M, m) -uniformly” extended to bounded, invertible operators in $\mathcal{L}(V_\sigma(\Omega), [V_\sigma(\Omega)]')$ that fulfill further “ (M, m) -uniform” resolvent bounds, have bounded inverses, well-defined fractional powers and fulfill $H_{\frac{1}{2}}(\tilde{\mathbb{T}}_{M,m}) = H_{\frac{1}{2}}(\tilde{\mathbb{T}}_{M,m}^*) = V_\sigma(\Omega)$ with “ (M, m) -uniform” norm equivalence.

The following proof of **requirement (ii) part (b)** of theorem 4.5.2 for the Navier-Stokes case that we are currently looking at will be very similar to the proof of that requirement for the convection-diffusion operator in the previous subsection. For completeness sake and because it is an important result, we will again go over the details. As usual, we set $t_m := \frac{mT}{M}$ for all $M \in \mathbb{N}$ and $m \in \{0, \dots, M-1\}$.

From various embeddings and part (ii) of theorem 3.8.2 we deduce for all $M \in \mathbb{N}$, $m \in \{0, \dots, M-1\}$, $v \in V_\sigma(\Omega)$ and $x := \tilde{\mathbb{A}}(t_m)^{-\frac{1}{2}}v \in H^2(\Omega)^d$ that

$$\begin{aligned} \|\tilde{\mathbb{T}}_{M,m}^{\frac{1}{2}}x\|_{H^1} &\leq C_{\mathbb{S}1}^{(0)}\|\tilde{\mathbb{T}}_{M,m}x\|_{L^2} \\ &\leq C_{\mathbb{S}1}^{(0)}(\|\nu\Delta x\|_{L^2} + \|(\hat{u}_{M,m} \cdot \nabla)x + (x \cdot \nabla)\hat{u}_{M,m}\|_{L^2} + |\tilde{c}|\|x\|_{L^2}) \\ &\leq C_{\mathbb{S}1}^{(1)}(\|\nu\Delta x\|_{L^2} + C_{\mathbb{T},H^2}\|x\|_{H^1} + |\tilde{c}|\|x\|_{L^2}) \leq C_{\mathbb{S}1}^{(2)}\|x\|_{H^2} \\ &\leq C_{\mathbb{S}1}^{(3)}\|\tilde{\mathbb{A}}(t_m)x\|_{L^2} \leq C_{\mathbb{S}1}^{(4)}\|\tilde{\mathbb{A}}^{\frac{1}{2}}(t_m)x\|_{H^1} = C_{\mathbb{S}1}^{(4)}\|v\|_{H^1}, \end{aligned}$$

where all constants — in particular the last one, $C_{\mathbb{S}1}^{(4)} > 0$ — are independent of v , m and M . Thus, the first part of requirement (ii) (b) is shown.

To prove the second and third part of requirement (ii) (b), we assume without loss of generality that $\tilde{c} = d$, i.e., in the operators $\tilde{\mathbb{T}}_{M,m} - \tilde{\mathbb{A}}(t_m)$ the zero-order term vanishes. Furthermore, we abbreviate the presentation by setting $u_{M,m} := \hat{u}_{M,m} - u(t_m)$ for all $M \in \mathbb{N}$ and $m \in \{0, \dots, M-1\}$.

Let $v \in V_\sigma(\Omega)$ and $y := \tilde{\mathbb{T}}_{M,m}^{-\frac{1}{2}}v \in H^2(\Omega)$. Since the $\tilde{\mathbb{T}}_{M,m}$ are sectorial operators with 0 in their respective sectors and “ (M, m) -uniform” resolvent bounds on those sectors, we have for all $M \in \mathbb{N}$, $m \in \{0, \dots, M-1\}$ that

$$\|y\|_{L^2}^2 \leq \|\tilde{C}_1 \tilde{\mathbb{T}}_{M,m} y\|_{L^2}^2 \quad (4.19)$$

with a constant $\tilde{C}_1 > 0$ independent of v , m and M .

It is now not hard to see that for the $\tilde{\mathbb{T}}_{M,m}$ one can get a “ (M, m) -uniform” H^2 -regularity result similar to the t -uniform Oseen H^2 -regularity (3.20) by using our assumption (4.18) and similar arguments as were used for the Oseen result. Together with (4.19), this provides us with the bound

$$\|y\|_{H^2}^2 \leq \|\tilde{C}_2 \tilde{\mathbb{T}}_{M,m} y\|_{L^2}^2$$

with a constant $\tilde{C}_2 > 0$ independent of v , m and M .

Using this bound, the product rule of differentiation and again a variety of different continuous Sobolev embeddings, we get for all $M \in \mathbb{N}$, $m \in \{0, \dots, M-1\}$ that

$$\begin{aligned} & \| (u_{M,m} \cdot \nabla) y + (y \cdot \nabla) u_{M,m} \|_{H^1}^2 \\ &= \| \nabla ((u_{M,m} \cdot \nabla) y + (y \cdot \nabla) u_{M,m}) \|_{L^2}^2 + \| (u_{M,m} \cdot \nabla) y + (y \cdot \nabla) u_{M,m} \|_{L^2}^2 \\ &\leq C_{\mathbb{S}2}^{(0)} (\| \nabla y \nabla u_{M,m} \|_{L^2}^2 + \| \nabla u_{M,m} \nabla y \|_{L^2}^2 + \| (u_{M,m} \cdot \nabla) \nabla y \|_{L^2}^2 \\ &\quad + \| (y \cdot \nabla) \nabla u_{M,m} \|_{L^2}^2 + \| (u_{M,m} \cdot \nabla) y \|_{L^2}^2 + \| (y \cdot \nabla) u_{M,m} \|_{L^2}^2) \\ &\leq C_{\mathbb{S}2}^{(1)} (\| \nabla y \|_{L^6}^2 \| \nabla u_{M,m} \|_{L^3}^2 + \| \nabla u_{M,m} \|_{L^3}^2 \| \nabla y \|_{L^6}^2 + \| u_{M,m} \|_{L^\infty}^2 \| y \|_{H^2}^2 \\ &\quad + \| y \|_{L^\infty}^2 \| u_{M,m} \|_{H^2}^2 + \| u_{M,m} \|_{L^\infty}^2 \| y \|_{H^1}^2 + \| y \|_{L^6}^2 \| \nabla u_{M,m} \|_{L^3}^2) \\ &\leq C_{\mathbb{S}2}^{(2)} (\| y \|_{H^2}^2 \| u_{M,m} \|_{\mathbb{S}_{\frac{3}{4}+\epsilon}}^2 + \| u_{M,m} \|_{\mathbb{S}_{\frac{3}{4}+\epsilon}}^2 \| y \|_{H^2}^2 + \| u_{M,m} \|_{\mathbb{S}_{\frac{3}{4}+\epsilon}}^2 \| y \|_{H^2}^2 \\ &\quad + \| y \|_{\mathbb{S}_{\frac{3}{4}+\epsilon}}^2 \| u_{M,m} \|_{H^2}^2 + \| u_{M,m} \|_{\mathbb{S}_{\frac{3}{4}+\epsilon}}^2 \| y \|_{H^1}^2 + \| y \|_{H^1}^2 \| u_{M,m} \|_{\mathbb{S}_{\frac{3}{4}+\epsilon}}^2) \\ &\leq C_{\mathbb{S}2}^{(3)} (\| y \|_{H^2}^2 \| u_{M,m} \|_{H^2}^2) \leq C_{\mathbb{S}2}^{(3)} (\| y \|_{H^2}^2 2(C_{\mathbb{T}, H^2}^2 + \| u \|_{C([0,T]; H^2(\Omega)^d)}^2)) \\ &\leq C_{\mathbb{S}2}^{(4)} \| \tilde{\mathbb{T}}_{M,m} y \|_{L^2}^2 \leq C_{\mathbb{S}2}^{(5)} \| \tilde{\mathbb{T}}_{M,m}^{\frac{1}{2}} y \|_{H^1}^2 = C_{\mathbb{S}2}^{(5)} \| v \|_{H^1}^2 \end{aligned}$$

with some $\epsilon > 0$. We used this ϵ here to emphasize that in almost all steps of proving the above inequality, we only need the regularity $u \in C([0, T]; H_{\frac{3}{4}+\epsilon}(\mathbb{S}))$ and $\{\hat{u}_{M,m} : M \in \mathbb{N}, m \in \{0, \dots, M-1\}\} \subset H_{\frac{3}{4}+\epsilon}(\mathbb{S})$. At precisely one point we do need the stronger H^2 -regularity for both the exact solution and the $\hat{u}_{M,m}$. Unfortunately, we do not see a way to get by without it — even for larger β . So this is where we really do need our assumption (4.18) for the $\hat{u}_{M,m}$. Notice, however, that in order to fulfill requirement (i) of theorem 4.5.2, the exact solution was already required to have the regularity $u \in H^1(0, T; H^2(\Omega)^d \cap V^\sigma(\Omega))$ which implies $u \in C([0, T]; H^2(\Omega)^d)$ because of the Bochner space embeddings listed in theorem 3.1.2.

Once again, all constants that arise in the proof of the above inequality — in particular the last one, $C_{\mathbb{S}2}^{(5)} > 0$ — are independent of v , m and M , thus showing the second part of requirement (ii) (b).

For the third and final part of requirement (ii) (b), we use the fact that for all $x_1, x_2, x_3 \in V_\sigma(\Omega)$ integration by parts gives us:

$$((x_1 \cdot \nabla) x_2, x_3)_{L^2} = -(\operatorname{div}(x_1), \bar{x}_2 \cdot x_3)_{L^2} - ((x_1 \cdot \nabla) \bar{x}_3, \bar{x}_2)_{L^2} = -((x_1 \cdot \nabla) \bar{x}_3, \bar{x}_2)_{L^2}.$$

Using this, we get for all $z, w \in V_\sigma(\Omega)$ and all $M \in \mathbb{N}$, $m \in \{0, \dots, M-1\}$ that:

$$\begin{aligned}
 |\langle (u_{M,m} \cdot \nabla)z + (z \cdot \nabla)u_{M,m}, w \rangle| &= |((u_{M,m} \cdot \nabla)z + (z \cdot \nabla)u_{M,m}, w)_{L^2}| \\
 &= |-(u_{M,m} \cdot \nabla)\bar{w}, \bar{z})_{L^2} + (z \cdot \nabla)u_{M,m}, w)_{L^2}| \\
 &\leq \|u_{M,m}\|_{L^\infty} \|\nabla w\|_{L^2} \|z\|_{L^2} + \|z\|_{L^2} \|\nabla u_{M,m}\|_{L^3} \|w\|_{L^6} \\
 &\leq C_{\mathbb{S}3}^{(0)} (C_{\mathbb{T}, H^2} + \|u\|_{C([0,T]; H^2(\Omega)^d)}) \|w\|_{H^1} \|z\|_{L^2} \\
 &= C_{\mathbb{S}3}^{(1)} \|w\|_{H^1} \|z\|_{L^2},
 \end{aligned}$$

and utilizing $H_{\frac{1}{2}}(\tilde{\mathbb{T}}_{M,m}^*) = V_\sigma(\Omega)$ with “ (M, m) -uniform” norm equivalence we further obtain

$$\begin{aligned}
 \|z\|_{L^2} &= \sup_{\substack{h \in H_\sigma(\Omega) \\ h \neq 0}} \frac{|(z, h)_{L^2}|}{\|h\|_{L^2}} = \sup_{\substack{w \in V_\sigma(\Omega) \\ w \neq 0}} \frac{|(z, \tilde{\mathbb{T}}_{M,m}^{*\frac{1}{2}} w)_{L^2}|}{\|\tilde{\mathbb{T}}_{M,m}^{*\frac{1}{2}} w\|_{L^2}} \leq C_{\mathbb{S}3}^{(2)} \sup_{\substack{w \in V_\sigma(\Omega) \\ w \neq 0}} \frac{|(z, \tilde{\mathbb{T}}_{M,m}^{*\frac{1}{2}} w)_{L^2}|}{\|w\|_{H^1}} \\
 &= C_{\mathbb{S}3}^{(2)} \sup_{\substack{w \in V_\sigma(\Omega) \\ w \neq 0}} \frac{|(\tilde{\mathbb{T}}_{M,m}^{\frac{1}{2}} z, w)_{L^2}|}{\|w\|_{H^1}} = C_{\mathbb{S}3}^{(2)} \sup_{\substack{w \in V_\sigma(\Omega) \\ w \neq 0}} \frac{|\langle \tilde{\mathbb{T}}_{M,m}^{\frac{1}{2}} z, w \rangle|}{\|w\|_{H^1}} = C_{\mathbb{S}3}^{(2)} \|\tilde{\mathbb{T}}_{M,m}^{\frac{1}{2}} z\|_{V_\sigma'},
 \end{aligned}$$

where all constants — in particular the last one, $C_{\mathbb{S}3}^{(2)} > 0$ — are independent of z , w , m and M . Setting $z := \tilde{\mathbb{T}}_{M,m}^{-\frac{1}{2}} v \in H_0^1(\Omega)$ for all $v \in H_0^1(\Omega)$, $M \in \mathbb{N}$, $m \in \{0, \dots, M-1\}$ and using the two inequalities above, we obtain the third part of requirement (ii) (b).

Finally, we want to apply W-methods with inexact Jacobians to a convection diffusion operator that is coupled with simple monomials. Further above we already examined this problem setup for ROW methods. For our W-method variants here, we will, similarly to the Navier-Stokes case, calculate Fréchet derivatives at other vectors than the exact solution vectors. Since the arguments here are very similar to the ones we used previously for the incompressible Navier-Stokes Problem, we shorten the presentation and omit the details.

W-Methods for Some Reaction-Diffusion Problems

In problem 3.7.1 let $\mathbb{A} := \mathbb{D}$ with \mathbb{D} defined as in section 3.3 and pair it with a nonlinearity defined as $N(v) := v^r$ for some $r \in \mathbb{N}$ and all $v \in H_0^1(\Omega)$. We already saw above, in our examination of ROW methods for reaction-diffusion equations, that for $r \leq 3$ and sufficient regularity for the exact solution u of problem 3.7.1, all requirements of theorem 4.5.1 are then fulfilled. We will now define the approximate Fréchet derivatives for our W-methods in a similar way as we did for the semilinear Navier-Stokes problem.

Even though the “operator-errors” that we have in mind only contain zero-order terms, we will work with some $\beta > 0$. We do this because in order to fulfill (4.12) with $\beta = 0$, we would need more than H^1 -regularity for the exact solution anyway. So having $\beta \neq 0$ does not demand too much regularity via requirement (i) of theorem 4.5.2. This is true even for $r = 2$.

So let $\beta \in (\frac{1}{4}, 1]$. We demand $\beta > \frac{1}{4}$ here because it allows for a relatively easy analysis. For example **requirement (i)** of theorem 4.5.2 can then be shown using similar arguments as were used

in theorem 3.8.3 if the exact solution has the regularity $u \in H^1(0, T; (\mathbb{D}^{-\beta}(H_0^1(\Omega)), \|\mathbb{D}^\beta \cdot\|_{H^1}))$.

For all $M \in \mathbb{N}$, $m \in \{0, \dots, M-1\}$ and some $\hat{u}_{M,m} \in H_0^1(\Omega)$ we set

$$\mathbb{T}_{M,m}v := \mathbb{D}v - r\hat{u}_{M,m}^{r-1}v$$

for all $v \in H_0^1(\Omega)$. Similarly to the Navier-Stokes case, this includes the choice of only sporadically updating the Jacobian but also requires us to assume

$$\exists C_{\mathbb{D},H} > 0 \forall M \in \mathbb{N} \forall m \in \{0, \dots, M-1\} : \|\mathbb{D}^\beta \hat{u}_{M,m}\|_{H^1} \leq C_{\mathbb{D},H},$$

in order to show that **requirement (ii)** of theorem 4.5.2 is fulfilled.

We will not prove the following statement in detail here. It can be shown with techniques that are very similar to the ones we used directly above in the two subsections on W-methods for the linear case and the Navier-Stokes case. Elliptic regularity and Sobolev embeddings again play a major role, of course. We have:

For $u \in H^1(0, T; (\mathbb{D}^{-\beta}(H_0^1(\Omega)), \|\mathbb{D}^\beta \cdot\|_{H^1}))$ requirements (i) and (ii) in theorem 4.5.1 and requirement (i) in theorem 4.5.2 are fulfilled for any $r \leq 3$. Furthermore, with the above described definitions and assumptions for the $\mathbb{T}_{M,m}$ and $\hat{u}_{M,m}$, requirement (ii) of theorem 4.5.2 is also fulfilled for any $r \leq 3$.

As was the case for Navier-Stokes, a larger value for β unfortunately does not help to reduce the regularity requirements on the exact solution or the $\hat{u}_{M,m}$.

4.5.2 Smoothness Assumptions and Order Reduction

We want to briefly touch on the issue that the proofs of high order error estimates need also high temporal regularity of the exact solution of a given problem. Roughly speaking, the error estimates are based on Taylor polynomials of the exact solution, so suitable regularity — depending on the degree of the polynomial — is needed.

Furthermore, the constants in the error estimates depend on the exact solution and its derivatives. These constants might become larger when different data leads to different exact solutions that are still *sufficiently* smooth, but in a way less so — for example through growing oscillations. And then there are cases where the data is so irregular that the exact solution has temporal derivatives with singularities at 0 and the above introduced error estimates are not applicable anymore. At the end of this section, we briefly mention other — lower order — error estimates that still work for non-smooth solutions.

There are three main types of problem data that influence the smoothness of a solution: the initial condition, the forcing term and the boundary data. Additionally, some conditions on the compatibility of the initial condition and the forcing term with the boundary data also play an important role.

As an example, we look at the initial value problem for a simple semilinear equation

$$\begin{aligned} u_t(t) + \mathbb{A}u(t) &= 0 \quad \text{for all } t \in [0, T], \\ u(0) &= u_0 \end{aligned}$$

with a sectorial operator \mathbb{A} on some Hilbert space H and an initial condition u_0 . Then these compatibility conditions with the boundary data can be expressed as requiring $u_0 \in D(\mathbb{A}^p)$ for some $p \in \mathbb{N}$ that is related to the order of the error estimate. If \mathbb{A} is a differential operator with homogeneous dirchlet boundary conditions, this requirement means that not only u_0 itself needs to be zero on the boundary but also some of its derivatives.

These requirements are often called “unnatural” in the literature. To see that they really are problematic for some practical applications, consider for example a reaction-diffusion problem with different chemical substances that are initially separated. In the beginning, the chemical concentrations are then not even differentiable in space. Still, these applications do have solutions — albeit less temporally-smooth ones. In those cases, we need to fall back on non-smooth error estimates.

To show how a non-zero forcing term might also influence these compatibility conditions, we will look at a fairly simple example in more detail.

Problem 4.5.1 (Initial Boundary Value Problem for the Heat Equation). *Let $\Omega \subset \mathbb{R}^d$ be open and bounded for some $d \in \{2, 3\}$ and let $T \in \mathbb{R}_{>0}$. Then we seek a function $u : [0, T] \times \Omega \rightarrow \mathbb{R}$ so that*

$$\begin{aligned} \frac{\partial u}{\partial t}(t, x) - \Delta u(t, x) &= f(t, x) && \text{for all } (t, x) \in (0, T] \times \Omega, \\ u(t, x) &= 0 && \text{for all } (t, x) \in [0, T] \times \partial\Omega, \\ u(0, x) &= u_0(x) && \text{for all } x \in \Omega, \end{aligned}$$

where $u_0 : \Omega \rightarrow \mathbb{R}$ is the initial condition and $f : [0, T] \times \Omega \rightarrow \mathbb{R}$ is the forcing term.

It is well-known that this problem has a unique solution if u_0 and f are sufficiently smooth. Furthermore, the solution can be shown to have higher regularity if u_0 and f are sufficiently regular **and** fulfill certain compatibility conditions. A proof of the following result can be found for example in sections 7.1.2. and 7.1.3. of [16].

Theorem 4.5.3. *Let $m \in \mathbb{N} \cup \{0\}$. In problem 4.5.1*

(i) *let $u_0 \in H^{2m+1}(\Omega, \mathbb{R})$,*

(ii) *define $\tilde{f} : [0, T] \rightarrow \Omega^{\mathbb{R}}$ as $[\tilde{f}(t)](x) = f(t, x)$ for all $(t, x) \in [0, T] \times \Omega$ and let $\frac{d^k \tilde{f}}{dt^k} \in L^2(0, T; H^{2m-2k}(\Omega, \mathbb{R}))$ for all $k \in \{0, \dots, m\}$,*

(iii) *let $g_0 := u_0 \in H_0^1(\Omega, \mathbb{R})$ and for all $k \in \{1, \dots, m\}$ let $g_k := \frac{d^{k-1} \tilde{f}}{dt^{k-1}}(0) - \Delta g_{k-1} \in H_0^1(\Omega, \mathbb{R})$.*

Then the so specified problem 4.5.1 has a unique solution $u : [0, T] \times \Omega \rightarrow \mathbb{R}$ with the regularity

$$\frac{d^k \tilde{u}}{dt^k} \in L^2(0, T; H^{2m+2-2k}(\Omega, \mathbb{R})) \quad \text{for all } k \in \{0, \dots, m+1\},$$

where $\tilde{u} : [0, T] \rightarrow \Omega^{\mathbb{R}}$ is defined as $[\tilde{u}(t)](x) = u(t, x)$ for all $(t, x) \in [0, T] \times \Omega$.

In order to provide a solution with higher regularity, the above theorem clearly demands in requirement (iii) that higher derivatives of the initial condition and the right-hand side need to match up with the homogeneous dirichlet conditions in a certain way.

In section 6.2. of [69] the author proves a comparable result for the incompressible Navier-Stokes equations and gives necessary and sufficient conditions for the existence of solutions that are smooth up to $t = 0$. Among those conditions are compatibility conditions that are very similar to the ones given above for the heat equation.

The obvious question now is, why error estimates, which depend on very smooth solutions, are of much interest, when in applications the requirements for the existence of these smooth solutions are often not fulfilled. The answer lies in a property of parabolic equations that is commonly called “parabolic smoothing” and can be explained as follows:

Many parabolic problems with non-smooth data still have unique solutions — although the solutions are not very regular in time because some of their temporal derivatives have singularities at $t = 0$ and are thus not in $L^2(0, T; L^2(\Omega, \mathbb{R}))$. However, after their initial singularity, these temporal derivatives often have more regularity in the sense that they are in $L^2(c, T; H^m(\Omega, \mathbb{R}))$ for any $c \in (0, T)$ and for an $m \in \mathbb{N}$ that depends on the forcing term but not on the smoothness of the initial condition or any compatibility conditions. An instructive paper on the topic of parabolic smoothing is [47].

When solving a parabolic problem which has an exact solution with the above described low regularity near $t = 0$ but higher regularity elsewhere, a common approach is to resolve the initial transient phase with very small time steps to avoid losing too much accuracy despite a lower order of convergence. After that transient phase, usually higher order of convergence can be obtained and thus larger time steps may be used again to save computing time.

Nevertheless, to obtain a complete mathematical foundation even for non-smooth solutions, some authors have also developed semi-discrete error estimates with less smoothness requirements. First of all, these estimates show that numerical solvers are still applicable — albeit yielding lower order of convergence — if the exact solution is not smooth. Secondly, these estimates are also important for examining attractors of dynamical systems — see [39] and [46] for details. In the latter of those papers, the authors examine the discretization of SPEs by W-methods that approximate the exact Fréchet derivatives with the sectorial operator of the equation — a common situation that we also looked at in our examples above. They show that, in general, a W-method applied in this way essentially achieves only convergence of order one for non-smooth initial data.

4.6 Specific Rosenbrock-type Methods with Parameter Tables

To close this chapter, we present in detail the coefficients of some methods for the time discretization of our problems. As explained in the previous sections, there are many requirements that we want our methods to meet, not all of which can be fulfilled simultaneously — especially when we want to keep the implementation simple, i.e., the number of internal stages low.

The following methods are always assumed to be of the form 4.2.2 and will be given using the corresponding notation.

Numerical Method 4.6.1 (ROS2).

$s = 2$	$b_1 = \frac{1}{2}$
$\gamma = 1 \pm \frac{1}{\sqrt{2}}$	$b_2 = \frac{1}{2}$
$\alpha_{21} = 1$	$\gamma_{21} = -2\gamma$

A detailed description of the method ROS2 can be found for example in [70] — the publication in which the method was first introduced. Both choices for γ lead to an L-stable method, which has classical convergence order 2 when used as ROW method with exact Jacobians and also when it is used as a W-method with approximate Jacobians. In section 3.2. of [70] it is argued, though, that the choice $\gamma = 1 + \frac{1}{\sqrt{2}}$ leads to a more stable scheme for nonlinear problems while the choice $\gamma = 1 - \frac{1}{\sqrt{2}}$ leads to smaller error constants. At the end of section 4.3, we also make an argument as to why the choice $\gamma = 1 + \frac{1}{\sqrt{2}}$ could provide more stability compared to the other choice for γ when ROS2 is used as a W-method, i.e., with approximate Jacobians.

ROS2 does not fulfill any of the conditions in theorem 4.4.1 or theorem 4.4.2. Nonetheless, we use this well-known and frequently used method to gauge the quality of our other methods.

Numerical Method 4.6.2 (Scholz45).

$s = 2$	$b_1 = \frac{1}{2}$
$\gamma = \frac{1}{2}$	$b_2 = \frac{1}{2}$
$\alpha_{21} = 1$	$\gamma_{21} = -1$

We obtained the method Scholz45 from (4.5) of the paper [58]. The method has 2 stages, is A-stable and has classical convergence order 2 when used as ROW method with exact Jacobians. It also has classical convergence order 2 when used as W-method with approximate Jacobians. Scholz45 fulfills the conditions of theorem 4.4.2 and all of the conditions in theorem 4.4.1 up to $p = 2$. Its main weakness is that its stability function $g(z)$ tends to -1 as $\text{Re } z$ tends to $-\infty$, i.e., the method is not strongly A-stable or even strongly $A(\alpha)$ -stable for any $\alpha \in (0, \frac{\pi}{2})$. Nevertheless, the method Scholz45 with its very simple set of coefficients displays very promising results in the numerical experiments.

Numerical Method 4.6.3 (ROS3PW).

$s = 3$	$b_1 = 0.10566243270259355$
$\gamma = 0.78867513459481287$	$b_2 = 0.049038105676657971$
	$b_3 = 0.84529946162074843$
$\alpha_{21} = 1.5773502691896257$	$\gamma_{21} = -\alpha_{21}$
$\alpha_{31} = 0.5$	$\gamma_{31} = -0.67075317547305480$
$\alpha_{32} = 0$	$\gamma_{32} = -0.17075317547305482$

We got the method ROS3PW from section 3.3 of [52]. The method has 3 stages, is strongly A-stable and has classical convergence order 3 when used as ROW method with exact Jacobians. It has classical convergence order 2 when used as W-method with approximate Jacobians. ROS3PW fulfills all of the conditions in theorem 4.4.1 up to $p = 2$. In addition, the method fulfills the conditions of theorem 4.4.2.

Numerical Method 4.6.4 (ROS34PRW).

$s = 4$	$b_1 = 0.33303742833830591$
$\gamma = 0.435866521508459$	$b_2 = 0.71793326075422947$
	$b_3 = -0.48683721060099439$
	$b_4 = \gamma$
$\alpha_{21} = 0.87173304301691801$	$\gamma_{21} = -\alpha_{21}$
$\alpha_{31} = 1.4722022879435914$	$\gamma_{31} = -1.2855347382089872$
$\alpha_{32} = -0.31840250568090289$	$\gamma_{32} = 0.50507005541550687$
$\alpha_{41} = 0.81505192016694938$	$\gamma_{41} = -0.48201449182864348$
$\alpha_{42} = 0.5$	$\gamma_{42} = 0.21793326075422950$
$\alpha_{43} = -0.31505192016694938$	$\gamma_{43} = -0.17178529043404503$

We obtained the method ROS34PRW from section 4.2 of [49]. The method has 4 stages, is L-stable and has classical convergence order 3 when used as ROW method with exact Jacobians. It also has classical convergence order 3 when used as W-method with approximate Jacobians. ROS34PRW fulfills all of the conditions in theorem 4.4.1 up to $p = 2$ and fulfills *some* of the conditions in that theorem for $p = 3$ — see pages 52 and 53 of [51] for details. In addition, the method fulfills the conditions of theorem 4.4.2.

5 Discretization in Space by the Finite Element Method

We now give a brief introduction to the popular finite element method (FEM). In this presentation we restrict ourselves to variants that are suitable for our applications. Furthermore, because the finite element method is not the main focus of this work and to keep this chapter easy to read, we choose a more condensed and less formal style for our presentation here. For more details on the finite element method we refer the reader to one of the many available textbooks, such as [11] or [31] — with the latter of the two being specialized towards incompressible flow problems.

When using a Rosenbrock-type method in a fully-discrete algorithm to discretize in time an SPE of the form 3.7.1, we obtain, on a triplet of spaces (V, H, V') , a series of stationary problems, to which we then seek approximate solutions. All of these stationary problems are of the following form: Seek a $u \in V$ so that

$$\left(\gamma \mathbb{T}_m + \frac{1}{\tau_m} \right) u = f_{mi}, \quad (5.1)$$

where $m \in \mathbb{N}$ is the index of the current time step, $s \in \mathbb{N}$ the number of stages of the method, $i \in \{1, \dots, s\}$ the current stage, $\gamma > 0$ a moderately sized parameter of the method, τ_m the current time step size and $\mathbb{T}_m \in \mathcal{L}(V, V')$ the currently used linearization of the original equation. The right-hand side $f_{mi} \in V'$ can be obtained explicitly from fully-discrete solutions from the previous time step and previous stages.

In accordance with our usual framework — and covering all our numerical experiments with the applications from section 4.5.1 — we assume that \mathbb{T}_m can be viewed as a sectorial operator on H that fulfills $H_{\frac{1}{2}}(\mathbb{T}_m) = V$ and can be extended to an operator in $\mathcal{L}(V, V')$. Notice that if the time steps are small enough, the operator $\gamma \mathbb{T}_m + \frac{1}{\tau_m}$ is invertible, i.e., (5.1) has a unique solution.

Since most of the literature on the finite element method uses formulations of the given problems that involve bilinear or sesquilinear forms, we will do so as well in order to avoid unnecessary technical difficulties. The connection to our framework of SPEs and sectorial operators is as follows:

For some $u \in V$, the formulation (5.1) is equivalent to the variational formulation

$$a(u, v) = \langle f, v \rangle \quad \text{for all } v \in V, \quad (5.2)$$

if we set $f = f_{mi}$ and for all $u, v \in V$ the sesquilinear form $a : V \times V \rightarrow \mathbb{C}$, fulfills

$$a(u, v) = (\tilde{\mathbb{T}}_m^{\frac{1}{2}} u, \tilde{\mathbb{T}}_m^{*\frac{1}{2}} v)_H - d(u, v)_H \left[= \langle \gamma \mathbb{T}_m u, v \rangle + \frac{1}{\tau_m} (u, v)_H \right],$$

with $\tilde{\mathbb{T}}_m := \gamma \mathbb{T}_m + \frac{1}{\tau_m} + d$ and a $d \in \mathbb{R}$ large enough, so that $0 \in \rho(\tilde{\mathbb{T}}_m)$. As already mentioned, if the time steps are small enough, $\gamma \mathbb{T}_m + \frac{1}{\tau_m}$ is already invertible — in that case we can choose $d = 0$.

In practice, a is, of course, not obtained through fractional powers of a given operator, but by multiplication of a strong PDE formulation with test functions, integration over the domain and suitable integration by parts to reduce regularity requirements on the solution and the test functions.

Usually, this way of constructing the bilinear form does not involve complex-valued spaces. Furthermore, spectral properties are generally not of much interest in this variational framework. Hence, we assume from now on that all spaces are *real* vector spaces and that a is a bilinear form that maps into \mathbb{R} . Thus, in contrast to the previous chapters, we use for some domain $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, the notation $L^2(\Omega) := L^2(\Omega, \mathbb{R})$, $H_0^1(\Omega) := H_0^1(\Omega, \mathbb{R})$ and so on — omitting the field \mathbb{R} .

The idea now is to look for approximate solutions of (5.2) in a finite dimensional subspace V_h of V . The standard situation is the following: V is a Sobolev space and the space V_h consists of piecewise polynomials defined on a mesh, which covers or lies within a given domain $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, and is made up of polytopes whose size is in some way described by the mesh parameter $h > 0$. Ambiguously, both the polytopes *and* the piecewise polynomials are often referred to as the “elements”. In order to fulfill our requirement that V_h is a proper subspace of V (in this case the method is called a conforming finite element method), further conditions in regard to the domain, the mesh and global properties of the piecewise polynomials — such as continuity across polytopes — have to be fulfilled.

The strategy for finding an approximation to the exact solution of (5.2) is to seek a $u_h \in V_h$ so that

$$a_h(u_h, v_h) = \langle f_h, v_h \rangle \quad \text{for all } v_h \in V_h. \quad (5.3)$$

Here, the bilinear form $a_h : V_h \times V_h \rightarrow \mathbb{R}$ and the right-hand side $f_h \in V_h'$ can either just be the restrictions $a_h = a|_{V_h \times V_h}$ and $f_h = f|_{V_h}$ or modifications thereof — in the latter case, the method is usually called a stabilized finite element method.

If $\{\phi_{hj} \in V_h : 1 \leq j \leq s\}$ is a basis of V_h and $\tilde{u}_h \in \mathbb{R}^s$, then $u_h := \sum_{j=1}^s (\tilde{u}_h)_j \phi_{hj} \in V_h$ solves (5.3) if and only if \tilde{u}_h solves the linear system

$$A \tilde{u}_h = \tilde{f}_h, \quad (5.4)$$

where

$$A_{ij} = a_h(\phi_j, \phi_i) \quad \text{for all } i, j \in \{1, \dots, s\} \quad \text{and} \quad (\tilde{f}_h)_j = \langle f_h, \phi_{hj} \rangle \quad \text{for all } j \in \{1, \dots, s\}.$$

The vector \tilde{f}_h can in many cases only be approximated — depending on the data f . The so-called stiffness matrix $A \in \mathbb{R}^{s \times s}$, on the other hand, is usually (not always !) determined *exactly* from a given basis and the bilinear form a_h . Its assembly is often numerically costly and one of the driving factors for seeking out time discretizations of nonlinear parabolic PDEs that reduce the number of times this matrix has to be rebuilt — such as Rosenbrock-type methods.

Considering the above way to treat a given variational problem, the following questions come to mind immediately:

1. Which properties of a and f are sufficient/necessary for the existence and uniqueness of a solution to the continuous variational formulation (5.2)?
2. Which properties of a_h , f and the spaces $\{V_h \leq V : h \in (0, 1)\}$ are sufficient/necessary for the existence and uniqueness of a solution to the discrete variational formulation (5.3)?
3. If both the discrete and the continuous variational formulation have unique solutions, what is known about the difference of the two solutions in various norms, and how do the parameter h and the properties of the spaces $\{V_h \leq V : h \in (0, 1)\}$ influence these errors?

We will not attempt to give answers to these questions for the general formulations above. Rather, in the next sections, we will look at specific applications of the finite element method that provide the spatial discretization for our numerical experiments. Since these ways of using the finite element method have been extensively covered in the literature, we do not go into the technical details.

Notice that one could address the first question by returning to the framework of sectorial operators again. In particular, we saw at the beginning of this chapter that unique solvability of (5.1) is directly given if the time steps are small enough. In the following sections however, we will continue our presentation with variational arguments. Unsurprisingly, the reactive term provided by the temporal discretization will also play an important role when examining unique solvability of both the continuous and the discrete variational formulations.

5.1 P_r - and Q_r -Elements

In this section we introduce the meshes that will be used for the remainder of this chapter. To keep this presentation simple, we assume that $\Omega \subset \mathbb{R}^2$ is a convex polygon. We want to mention, though, that the following arguments and results can be extended to the three dimensional case and that the finite element method can — with some adjustments — also be applied to domains that are not convex or have curved boundaries.

Now we further assume that there exists a mesh width $h \in \mathbb{R}_{>0}$, a mesh $\mathcal{T}_h \subset \mathcal{P}(\mathbb{R}^2)$ and an anisotropy parameter $\kappa > 0$ so that the following requirements are fulfilled:

- (i) The elements $T \in \mathcal{T}_h$ are either all triangles or all convex quadrilaterals.
- (ii) All elements $T \in \mathcal{T}_h$ are closed, i.e., they include their boundaries. Furthermore, the intersection of any two elements is either empty or a common vertex of the two elements or a common edge. Meshes with this property are called regular.
- (iii) The elements $T \in \mathcal{T}_h$ cover the domain Ω exactly, i.e., $\overline{\Omega} = \bigcup_{T \in \mathcal{T}_h} T$.
- (iv) For any element $T \in \mathcal{T}_h$ we denote by h_T the diameter of the smallest closed circle that contains T and by ρ_T the diameter of the largest closed circle within T . We now assume that

- (a) for all $T \in \mathcal{T}_h$ we have $h_T \leq \kappa \rho_T$ (if each mesh in a sequence of meshes fulfills this with the same κ , then the sequence of meshes is called quasi-uniform),
- (b) for all $T \in \mathcal{T}_h$ we have $\max_{T \in \mathcal{T}_h} h_T \leq \kappa \rho_T$, (if each mesh in a sequence of meshes fulfills this with the same κ , then the sequence of meshes is called uniform — note that uniformity is stronger than quasi-uniformity),
- (c) we have $h = \max_{T \in \mathcal{T}_h} h_T$, i.e., the parameter h describes the size of the largest element of the mesh \mathcal{T}_h .

Based on this mesh, we can now define discrete spaces. To do this, we first introduce spaces of polynomials.

For any $r \in \mathbb{N}_0$ we define P_r to be the set of all functions $u : M \rightarrow \mathbb{R}$ which are defined on some subset $M \subset \mathbb{R}^2$ and have the form

$$u(x, y) = \sum_{\substack{0 \leq i+j \leq r \\ 0 \leq i, j}} a_{ij} x^i y^j$$

for all $(x, y) \in M$ and some coefficients $a_{ij} \in \mathbb{R}$, $i, j \in \{0, \dots, r\}$, $0 \leq i + j \leq r$.

For any $r \in \mathbb{N}_0$ we define Q_r to be the set of all functions $u : \mathbb{R}^2 \rightarrow \mathbb{R}$ which are defined on some subset $M \subset \mathbb{R}^2$ and have the form

$$u(x, y) = \sum_{0 \leq i, j \leq r} a_{ij} x^i y^j$$

for all $(x, y) \in M$ and some coefficients $a_{ij} \in \mathbb{R}$, $i, j \in \{0, \dots, r\}$.

For any $r \in \mathbb{N}_0$ we now define the discrete spaces to be used in our finite element method as follows:

$$P_r(\mathcal{T}_h) := \{u \in C(\overline{\Omega}) : u|_T \in P_r \text{ for all } T \in \mathcal{T}_h\}$$

and

$$Q_r(\mathcal{T}_h) := \{u \in C(\overline{\Omega}) : u|_T \in Q_r \text{ for all } T \in \mathcal{T}_h\}.$$

Based on the above listed requirements on the mesh, several important properties of the discrete spaces can be shown. For example we have

$$P_r(\mathcal{T}_h), Q_r(\mathcal{T}_h) \subset H^1(\Omega), \tag{5.5}$$

i.e., these spaces are so-called H^1 -conform finite element spaces.

5.1.1 P_r - and Q_r -Elements for the Stationary Convection-Diffusion Equation

When using Rosenbrock-type methods in a fully-discrete algorithm to discretize in time a convection-diffusion problem with sufficiently regular forcing term and initial condition, the resulting stationary problems at the m -th time step can, in accordance with (5.1) and (5.2), be written in

the following variational formulation: Seek a $u \in V := H_0^1(\Omega)$ so that for all $v \in V$ we have

$$a(u, v) := \nu(\nabla u, \nabla v)_{L^2} + ((b \cdot \nabla)u, v)_{L^2} + (cu, v)_{L^2} + \frac{1}{\tau_m}(u, v)_{L^2} = (f, v)_{L^2} \quad (5.6)$$

where $\nu > 0$, $b \in L^\infty(\Omega)^2$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$ and the current time step size $\tau_m > 0$. In general, f depends on fully-discrete solutions from previous time steps and stages. In section 4.5.1 we examined the semi-discrete application of ROW methods and specific W-methods which use the Laplacian as an approximation to the full convection-diffusion operator to convection-diffusion problems. The above formulation (5.6) covers all our numerical experiments corresponding to those applications.

For all $u \in V$ we get

$$\begin{aligned} a(u, u) &= \nu \|\nabla u\|_{L^2}^2 + ((b \cdot \nabla)u, u)_{L^2} + (cu, u)_{L^2} + \frac{1}{\tau_m} \|u\|_{L^2}^2 \\ &\geq \nu \|\nabla u\|_{L^2}^2 - \|b\|_{L^\infty} \|\nabla u\|_{L^2} \|u\|_{L^2} + \left(\frac{1}{\tau_m} - \|c\|_{L^\infty} \right) \|u\|_{L^2}^2 \\ &\geq \nu \|\nabla u\|_{L^2}^2 - \frac{\nu}{2} \|\nabla u\|_{L^2}^2 - \frac{\|b\|_{L^\infty}}{2\nu} \|u\|_{L^2}^2 + \left(\frac{1}{\tau_m} - \|c\|_{L^\infty} \right) \|u\|_{L^2}^2 \\ &= \frac{\nu}{2} \|\nabla u\|_{L^2}^2 + \left(\frac{1}{\tau_m} - \|c\|_{L^\infty} - \frac{\|b\|_{L^\infty}}{2\nu} \right) \|u\|_{L^2}^2, \end{aligned}$$

using Young's inequality in the second to last step. Hence, we see that a is coercive if the time steps are small enough. Since a is obviously bounded, we then get existence and uniqueness of the continuous variational formulation from the classic Lax-Milgram theorem. Elliptic regularity results (see theorem 3.3.2) also guarantee that the solution is in $H^2(\Omega)$.

For the higher regularity $b \in W^{1,\infty}(\Omega)^2$, one can use integration by parts to show the alternative coercivity bound

$$a(u, u) \geq \nu \|\nabla u\|_{L^2}^2 + \left(\frac{1}{\tau_m} + \text{ess inf}(c - \text{div } b) \right) \|u\|_{L^2}^2$$

for all $u \in V$. Here, we make use of the essential infimum, which for any $w \in L^\infty(\Omega)^d$, $d \in \mathbb{N}$, takes the value $\text{ess inf}(w) = \|w\|_{L^\infty} - \|w - \|w\|_{L^\infty}\|_{L^\infty}$. The advantage of this bound, compared to the previous one, is that the time step size, which is needed for unique solvability, does not depend on the possibly very small parameter ν here.

We now use the mesh and the spaces from the section above to introduce our discrete variational formulation. In that formulation, we look for approximate solutions to (5.6) in a space $V_h := V_h^r \in \{P_r(\mathcal{T}_h) \cap H_0^1(\Omega), Q_r(\mathcal{T}_h) \cap H_0^1(\Omega)\}$ with $r \in \{1, 2\}$. Notice how we suppress the index for the polynomial degree in order to increase readability. The corresponding bilinear form $a_h : V_h \times V_h \rightarrow \mathbb{R}$ and the forcing term $f_h \in L^2(\Omega)$ are defined by simply setting $a_h = a|_{V_h \times V_h}$ and $f_h = f$.

Since by (5.5) V_h is a finite dimensional subspace of V , it is itself a Hilbert space. Thus, we immediately get from the above coercivity bounds that the discrete variational formulation of seeking a $u_h \in V_h$ so that

$$a_h(u_h, v_h) = (f_h, v_h) \quad \text{for all } v_h \in V_h, \quad (5.7)$$

is uniquely solvable.

Now, for a time step size small enough for a to be coercive, let $u \in H^{r+1}(\Omega)$ be a solution to the continuous problem (5.6) and let $u_h \in V_h$ be a solution to the discrete problem (5.7). Then one can use the mesh properties introduced in section 5.1 to show the error estimate:

$$\|u - u_h\|_{H^1} \leq C_{FE,conv} h^r \|u\|_{H^{r+1}}, \quad (5.8)$$

with the constant $C_{FE,conv} > 0$ depending on κ (an upper bound for the anisotropy of all elements in the mesh — see section 5.1 for more information) and on ν, b, c, τ_m but not on the mesh size h . Also remember that $r \in \{1, 2\}$ denotes the polynomial degree of the finite element space.

We want to mention that it gets increasingly difficult to obtain reasonable numerical solutions to the stationary convection diffusion problem if ν becomes smaller in relation to $\|b\|_{L^\infty}$. The discrete solution will often have unnatural oscillations if the so-called element Péclet number $\frac{\|b\|_{L^\infty} h}{2\nu}$ is significantly larger than 1. This issue is commonly mitigated by using stabilized finite element methods, i.e., by not setting $a_h = a|_{V_h \times V_h}$ and $f_h = f$ but instead defining a_h/f_h as some other bilinear form/forcing term that provide solutions which still converge to the exact solution when h approaches zero but have less oscillations even for larger element Péclet numbers.

5.1.2 Q_r -Elements for an Oseen-Type Equation

Now we turn our attention to the Oseen-type equation that emerges from the application of Rosenbrock-type methods to the Navier-Stokes equations. We begin with the stationary problems that we obtain from using Rosenbrock-type methods in a fully-discrete algorithm to discretize in time the incompressible Navier-Stokes equations (see section 3.8 for details and notations).

At the m -th time step we are seeking a $v \in V_\sigma(\Omega)$ so that for all $w \in V_\sigma(\Omega)$ we have

$$\langle -\nu \Delta v + (\hat{u} \cdot \nabla) v + (v \cdot \nabla) \hat{u} + \frac{1}{\tau_m} v, w \rangle = \langle f, w \rangle \quad (5.9)$$

with $\nu > 0$, $f \in H^{-1}(\Omega)^2$, the current time step size $\tau_m > 0$ and the transport direction $\hat{u} \in V_\sigma(\Omega)$. In general, f and \hat{u} depend on fully-discrete solutions from previous time steps and stages. As usual, we employ the triplet of spaces identifications $H_0^1(\Omega)^2 \hookrightarrow L^2(\Omega)^2 \cong [L^2(\Omega)^2]' \hookrightarrow H^{-1}(\Omega)^2$, so in particular $v \cong (v, \cdot)_{L^2} \in H^{-1}(\Omega)^2$ holds for all $v \in H_0^1(\Omega)^2$.

In section 4.5.1 we examined the semi-discrete application of ROW methods and some W-method variants with inexact Fréchet derivatives to incompressible flow problems. If we assume the higher regularity $\hat{u} \in H^2(\Omega)^2 \cap V_\sigma(\Omega)$, the above formulation (5.9) covers all our numerical experiments corresponding to those incompressible flow applications from section 4.5.1.

Recently, a significant amount of research has been done on finite element methods that use discrete spaces which consist entirely of divergence free functions. To keep this presentation simple, however, and also to stay in line with our numerical experiments, we are not going to utilize completely divergence free discrete spaces for the following arguments. Therefore, we need to reformulate (5.9) on spaces that are *not* divergence free and we also need to include the pressure.

With arguments very similar to the ones used in the proof of part (i) of proposition 3.8.1, one can show that a function $v \in V_\sigma(\Omega) \subset V := H_0^1(\Omega)^2$ solves (5.9) if and only if there is a $p \in Q := Q(\Omega) = \{p \in L^2(\Omega) : (p, 1)_{L^2} = 0\}$ so that for all $(w, q) \in V \times Q =: X$ we have

$$\langle -\nu \Delta v + (\hat{u} \cdot \nabla) v + (v \cdot \nabla) \hat{u} + \text{grad } p + \frac{1}{\tau_m} v, w \rangle = \langle f, w \rangle, \quad (5.10)$$

$$\langle \text{div } v, q \rangle = 0. \quad (5.11)$$

If the data is regular enough so that $f \in L^2(\Omega)^2$, we can thus use the following equivalent variational formulation: We are seeking $(v, p) \in X$ so that

$$\begin{aligned} a((v, p), (w, q)) &:= \nu(\nabla v, \nabla w)_{L^2} + ((\hat{u} \cdot \nabla) v, w)_{L^2} + ((v \cdot \nabla) \hat{u}, w)_{L^2} \\ &\quad + \frac{1}{\tau_m} (v, w)_{L^2} - (p, \text{div } w)_{L^2} + (\text{div } v, q)_{L^2} \\ &= (f, w)_{L^2} \end{aligned} \quad (5.12)$$

for all $(w, q) \in X$. Unfortunately, this bilinear form a is not coercive on X , no matter how small the time steps are. We see this by plugging in an arbitrary $(v, p) \in X$ to get

$$a((v, p), (v, p)) = \nu \|\nabla v\|_{L^2}^2 + ((\hat{u} \cdot \nabla) v, v)_{L^2} + ((v \cdot \nabla) \hat{u}, v)_{L^2} + \frac{1}{\tau_m} \|v\|_{L^2}^2.$$

Since the right-hand side does not contain p , there will for any constant $C > 0$ and any $v \in V$ always exist a $p \in Q$ so that $C(\|v\|_{H^1}^2 + \|p\|_{L^2}^2)$ is not bounded from above by that right-hand side.

However, if for all $v, w \in V$ we set $a_1(v, w) := a((v, 0), (w, 0))$, we can once again use Sobolev embeddings and Young's inequality to prove for all $v \in V$ that

$$\begin{aligned} a_1(v, v) &\geq \nu \|\nabla v\|_{L^2}^2 - \|\nabla \hat{u}\|_{L^3} \|v\|_{L^6} \|v\|_{L^2} + \frac{1}{\tau_m} \|v\|_{L^2}^2 \\ &\geq \frac{\nu}{2} \|\nabla v\|_{L^2}^2 + \left(\frac{1}{\tau_m} - \frac{\|\hat{u}\|_{L^\infty}}{2\nu} \right) \|v\|_{L^2}^2 \end{aligned}$$

if $\hat{u} \in W^{1,3}(\Omega)^2$. With the higher regularity $\hat{u} \in W^{1,\infty}(\Omega)^2$, it is easy to acquire for all $v \in V$ the alternative bound

$$a_1(v, v) \geq \nu \|\nabla v\|_{L^2}^2 + \left(\frac{1}{\tau_m} - \|\nabla \hat{u}\|_{L^\infty} \right) \|v\|_{L^2}^2.$$

These bounds show that for small enough time steps, $a_1 : V \times V \rightarrow \mathbb{R}$ and $(a_1)|_{V_\sigma(\Omega) \times V_\sigma(\Omega)}$ are coercive bilinear forms. Similarly to the convection diffusion case, the advantage of the second bound is that the time step size, which is needed to make a_1 coercive, does not depend on the possibly small parameter ν there.

Now, for small enough time steps, the Lax-Milgram theorem guarantees the unique existence of a $v \in V_\sigma(\Omega)$ so that $a_1(v, w) = (f, w)_{L^2}$ for all $w \in V_\sigma(\Omega)$. Also, one can show that the so-called inf-sup condition

$$\inf_{p \in Q \setminus \{0\}} \sup_{w \in V \setminus \{0\}} \frac{(p, \text{div } w)_{L^2}}{\|p\|_{L^2} \|\nabla w\|_{L^2}} \geq \gamma \quad (5.13)$$

holds for some $\gamma > 0$. Furthermore, using theorem 3.4.3 as in remark 3.4.3 shows that there is a $p \in Q$ so that for all $w \in V$ we have

$$a_1(v, w) - (p, \operatorname{div} w)_{L^2} = (f, w)_{L^2}$$

and therefore

$$a((v, w), (p, q)) = a_1(v, w) + (\operatorname{div} v, q)_{L^2} - (p, \operatorname{div} w)_{L^2} = a_1(v, w) - (p, \operatorname{div} w)_{L^2} = (f, w)_{L^2}$$

for all $w \in V$ and $q \in Q$.

Hence, we have obtained a solution $(v, p) \in X$ to (5.12) if the time steps are small enough. Furthermore, it is not hard to see that from theorem 3.4.4 we even get the higher regularity $v \in H^2(\Omega)^2$ and $p \in H^1(\Omega)$.

As in the previous section on the convection diffusion equation, we now use the mesh and the spaces introduced at the beginning of section 5.1 to select the discrete spaces we want to utilize. In this case here, we restrict ourselves to quadrilateral elements, i.e., we choose $V_h := V_h^r := [Q_r(\mathcal{T}_h)^2 \cap H_0^1(\Omega)^2]$ and $Q_h := Q_h^r := Q_r(\mathcal{T}_h) \cap Q$ with $r \in \{1, 2\}$. Notice how we again suppress the index for the polynomial degree in order to increase readability. We immediately see that $Q_h \leq Q$, and from 5.5 we directly get $V_h \leq V$ as well.

However, for the definition of the discrete bilinear form, we cannot proceed as in the previous section, where the discrete bilinear form is just the restriction of the continuous one and coercivity is directly inherited to provide unique solvability also for the discrete variational formulation. The issue in the Oseen case here is the inf-sup condition (5.13), which holds for the spaces V , Q , but not necessarily if we replace V , Q with *any* pair of subspaces $U_V \leq V$, $U_Q \leq Q$. It is not difficult to see that losing the inf-sup condition for the discrete spaces means that one cannot obtain a *unique* discrete pressure solution anymore. And one can in fact show that a discrete inf-sup condition does *not* hold for our specific choice V_h/Q_h or any pair of the type P_r/P_r or Q_r/Q_r — so-called equal order elements. Even though there are other pairs of spaces of piecewise polynomials that *do* fulfill a discrete inf-sup condition, we use a different approach here.

A way to get by the inf-sup condition with equal order elements — and thus simplify the numerical implementation in many cases — is to utilize stabilization techniques. In our experiments, we opt for the popular local projection stabilization (LPS) that also helps to mitigate issues with convection dominated problems. It was first published by Becker and Braack — see [6]. The presentation here is predominantly based on the later papers [9] and [10], though.

The local projection stabilization is implemented by modifying the bilinear form in a certain way but keeping the right-hand side unchanged. The discrete variational problem of looking for $(v_h, p_h) \in X_h := V_h \times Q_h$ so that for all $(w_h, q_h) \in X_h$ we have

$$a_h((v_h, p_h), (w_h, q_h)) = (f, w_h)_{L^2} \tag{5.14}$$

is solved not for the choice $a_h = a|_{X_h}$ but for a different definition of a_h that requires the mesh to have a macrostructure. That is, we require that there exists a mesh $\mathcal{T}_{2h} \subset \mathcal{P}(\mathbb{R}^2)$ which has mesh size $2h = \max_{T \in \mathcal{T}_{2h}} h_T$, fulfills all requirements from section 5.1 and contains exactly four elements of the mesh \mathcal{T}_h . To define a_h we first introduce the space

$$D_{2h} := \{v \in L^2(\Omega)^d : v|_T^\circ \in Q_0 \text{ for all } T \in \mathcal{T}_{2h}\}$$

and a projection $\pi_h : L^2(\Omega) \rightarrow D_{2h}$ of which we require

$$(\pi_h v)|_T^\circ = \frac{1}{|T|} \int_T v \, d\Omega$$

for all $T \in \mathcal{T}_{2h}$.

Moreover, we need the so-called fluctuation operators

$$\begin{aligned} \kappa_{1,h} : L^2(\Omega) &\rightarrow L^2(\Omega), \quad v \mapsto \pi_h v - v \\ \text{and} \quad \kappa_{2,h} : L^2(\Omega)^2 &\rightarrow L^2(\Omega)^2, \quad \begin{pmatrix} v_x \\ v_y \end{pmatrix} \mapsto \begin{pmatrix} \kappa_{1,h}(v_x) \\ \kappa_{1,h}(v_y) \end{pmatrix}. \end{aligned}$$

Finally, we set $a_h : X_h \times X_h \rightarrow \mathbb{R}$ to $a_h := a|_{X_h} + s_h$ with $s_h : X_h \times X_h \rightarrow \mathbb{R}$ defined as

$$\begin{aligned} s_h((v_h, p_h), (w_h, q_h)) &= (\kappa_{2,h}((\hat{u} \cdot \nabla) v_h), \delta_h \kappa_{2,h}((\hat{u} \cdot \nabla) w_h))_{L^2} + (\kappa_{2,h}(\nabla p_h), \delta_h \kappa_{2,h}(\nabla q_h))_{L^2} \\ &\quad + (\kappa_{1,h}(\operatorname{div} v_h), \gamma_h \kappa_{1,h}(\operatorname{div} w_h))_{L^2} \end{aligned} \quad (5.15)$$

for all $((v_h, p_h), (w_h, q_h)) \in X_h \times X_h$ and some stabilization parameters $\delta_h, \gamma_h \geq 0$.

Now, if the time step size and the data are such that there exists $(v, p) \in H^{r+1}(\Omega)^2 \times H^{r+1}(\Omega)^2 \cap X$ which solves (5.12) and if the Péclet number $\frac{\|\hat{u}\|_{L^\infty h}}{2\nu}$ is larger than 1, then we get from the paper [10] that with values $\delta_h \sim h$ and $\gamma_h \sim h$ for the stabilization parameters, the problem of solving (5.14) with the discrete bilinear form defined as in (5.15) has a unique solution $(v_h, p_h) \in V_h \times Q_h = V_h^r \times Q_h^r$ which fulfills

$$\nu^{\frac{1}{2}} \|\nabla(v - v_h)\|_{L^2} + \|p - p_h\|_{L^2} \leq C_{FE, Oseen} h^r ((h + \nu)^{\frac{1}{2}} \|v\|_{H^{r+1}} + h^{\frac{1}{2}} \|p\|_{H^{r+1}}). \quad (5.16)$$

The constant $C_{FE, Oseen} > 0$ depends on κ (an upper bound for the anisotropy of all elements in the mesh — see section 5.1 for more information) and on \hat{u}, τ_m but not on ν or the mesh size h . Also remember that $r \in \{1, 2\}$ denotes the polynomial degree of the finite element space.

5.2 Notes on the Full Discretization

The complete analysis of a numerical algorithm for the solution of parabolic PDEs ideally contains estimates on the error between the exact solution and the fully-discrete — i.e., discrete in time *and* space — numerical solution. Even though we will not give detailed fully-discrete error estimates for our methods and applications, we want to at least provide a starting point for obtaining such estimates.

Let (V, H, V') be a triplet of spaces and let $u \in C([0, T]; H)$ with $u(t) \in V$, $\frac{du}{dt}(t) \in V'$ for all $t \in [0, T]$ be the exact solution of an initial value problem of the type 3.7.1. If that problem fulfills some additional requirements, we can apply the convergence theory from section 4.5 and use Rosenbrock-type methods with certain properties to construct a numerical solution $(u_m)_{m=0}^M \in V^{M+1}$, $M \in \mathbb{N}$ being the total number of time steps, which is discrete only in time and fulfills an error estimate of the type

$$\left(\tau \sum_{m=1}^M \|u_m - u(t_m)\|_V^2 \right)^{\frac{1}{2}} + \max_{0 \leq m \leq M} \|u_m - u(t_m)\|_H \leq C_{\text{Time}} \tau^q \quad (5.17)$$

where $\tau = \frac{T}{M}$, $t_m = m\tau$ for all $m \in \{0, \dots, M\}$, $q \in \{2, 3\}$ and $C_{\text{Time}} > 0$ is some constant independent of τ .

On the other hand, we saw in this chapter how to acquire discrete in space approximations $u_h \in V_h \leq V$ to solutions $\tilde{u} \in V$ of specific stationary problems. From (5.8) and (5.16) corresponding error estimates of the form

$$\|\tilde{u} - u_h\|_V \leq C_{\text{Space}} h \quad (5.18)$$

can be deduced, where $h > 0$ describes a finite element mesh size and the constant $C_{\text{Space}} > 0$ does not depend on h .

Unfortunately we cannot simply combine (5.17) and (5.18) via the triangle inequality to construct estimates on the error between the exact solution u and our fully-discrete approximate solution. This can be understood as follows:

Let $s \in \mathbb{N}$ be the number of stages in the time-stepping scheme. The fully-discrete approximation is obtained by successively seeking solutions $(\hat{k}_{mi,h})_{m=0}^M \in V_h^{M+1}$, $i \in \{1, \dots, s\}$, $m \in \{0, \dots, M-1\}$, to discrete stationary equations of the type

$$\left\langle \left(\hat{\mathbb{T}}_m + \frac{1}{\tau} \right) \hat{k}_{mi,h}, w_h \right\rangle = \langle \hat{f}_{mi}, w_h \rangle \quad \text{for all } w_h \in V_h, \quad (5.19)$$

where the operator $\hat{\mathbb{T}}_m \in \mathcal{L}(V, V')$ depends on fully-discrete solutions from previous time steps and the right-hand side $\hat{f}_{mi} \in V'$ depends on fully-discrete solutions from previous time steps and stages. On the other hand, the semi-discrete (in time) solution is calculated by successively seeking solutions $(k_{mi})_{m=0}^M \in V^{M+1}$, $i \in \{1, \dots, s\}$, $m \in \{0, \dots, M-1\}$, to continuous stationary equations of the type

$$\left\langle \left(\mathbb{T}_m + \frac{1}{\tau} \right) k_{mi}, w \right\rangle = \langle f_{mi}, w \rangle \quad \text{for all } w \in V. \quad (5.20)$$

where the operator $\mathbb{T}_m \in \mathcal{L}(V, V')$ depends on semi-discrete solutions from previous time steps and the right-hand side $f_{mi} \in V'$ depends on semi-discrete solutions from previous time steps and stages.

The issue here is that, in general, we introduce a spatial discretization error each time we solve a stationary problem of the type (5.19). And since, as we already said above, the $\hat{\mathbb{T}}_m$, \hat{f}_{mi} usually depend on fully-discrete solutions from previous time steps, whereas the \mathbb{T}_m , f_{mi} usually depend on *semi*-discrete solutions from previous time steps, we generally have $\hat{\mathbb{T}}_m \neq \mathbb{T}_m$ and $\hat{f}_{mi} \neq f_{mi}$. Therefore, we cannot directly combine the estimates and (5.17) and (5.18). One could say that the spatial errors introduced during the time-stepping procedure are also *propagated* by that procedure in a way that is not immediately obvious.

In most situations, it *is* possible, though, to decouple spatial and temporal errors. In chapter III of the book [37] Jens Lang presents such a decoupling through use of the splitting

$$u - u_{m,h} = (u - \Pi_h u) + (\Pi_h u - u_{m,h}) \quad \text{for all } m \in \{0, \dots, M\}, \quad (5.21)$$

where $\Pi_h : V \rightarrow V_h$ is a projection with

$$\|\Pi_h u\|_V \leq C_{\Pi_h, \text{cont}} \|u\|_V \quad (5.22)$$

$$\text{and } \|u - \Pi_h u\|_V \leq C_{\Pi_h, \text{approx}} \|u\|_V \quad (5.23)$$

for all $u \in V$ and some constants $C_{\Pi_h, cont}, C_{\Pi_h, approx} > 0$ independent of u and h . An example for such a projection in the case $V = H_0^1(\Omega) \not\hookrightarrow C(\Omega)$ for some convex polygon $\Omega \subset \mathbb{R}^2$ and $V_h = P_1(\mathcal{T}_h)$ for some triangular mesh \mathcal{T}_h with properties as in section 5.1, is the Scott-Zhang interpolation.

The first term in (5.21) is a spatial projection error that depends on the properties of Π_h , while the second term can be accessed through a spatially discretized and perturbed version of the original SPE. In chapter III of [37] Lang shows under the assumptions from theorem 4.5.1, the assumptions (5.22) and (5.23) on the spatial discretization and some additional assumptions, which are usually fulfilled in practice, that the following bound holds for the fully-discrete solution:

$$\begin{aligned} & \left(\tau \sum_{m=0}^M \|u_{m,h} - u(t_m)\|_V \right)^{\frac{1}{2}} + \max_{0 \leq m \leq M} \|u_{m,h} - u(t_m)\|_H \\ & \leq \sqrt{\tau} \|\Pi_h u(0) - u(0)\|_V + \|\Pi_h u(0) - u(0)\|_H + C_{\text{time,space}} (\tau^q + \|u - \Pi_h u\|_{H^2(0,T;V)}) \end{aligned} \quad (5.24)$$

with a constant $C_{\text{time,space}} > 0$ that does not depend on τ or h .

Notice how the comparison of this bound with the semi-discrete (in time) bound (5.17) shows that the semi-discrete estimate already gives an upper bound for the contribution of the temporal discretization error to the full error — for all sizes of h . In other words, on the right-hand side of the above bound (5.24), temporal and spatial discretization errors are decoupled. Thus, if the exact solution is spatially well approximated by Π_h (or even in V_h for all $t \in [0, T]$) so that $u(t) = \Pi_h u(t)$ on the whole of $[0, T]$, then one can expect to get a reliable upper bound for the fully-discrete error by only considering the semi-discrete error estimates.

That is exactly what we will be doing in chapter 6: we will choose a fine enough spatial discretization — and usually even exact solutions from the finite element spaces — so that we only observe the temporal error. Keep in mind, though, that we can never observe the true semi-discrete error and that the semi-discrete error estimates are not necessarily *sharp* bounds for the fully-discrete errors. If the spatial discretization is not fine enough (or the temporal discretization is too fine), the numerical methods might not suffer from stability issues or order reduction phenomena so that the semi-discrete error estimates could overestimate the error and a higher order than they predict might be observed. We will come back to this point in chapter 6.

We want to add that even though we did not work out the details, it looks as if Lang's method for constructing fully-discrete error estimates from a semi-discrete error estimate for ROW methods and spatial discretization techniques which fulfill certain requirements should also be possible with the semi-discrete *W-method* error estimate 4.5.2.

6 Numerical Experiments

Here we put the Rosenbrock-type time-stepping methods from section 4.6 to the test. We use Rothe’s method as explained in section 4.1, i.e., we first discretize in time and then in space. For the spatial discretization, we use the finite element method introduced in the previous chapter 5. The specific problems are implemented with the finite element toolkit Gascoigne 3D (see [20]) and are chosen to suit the physical applications of mass and heat transfer and incompressible flow from chapter 2.

Of course we would also like to examine the semi-discrete in time error analysis from section 4.5 in regard to those problems. Since there was little to no attention paid to the actual numeric value of any constants that appear in error estimates or step size restrictions, the numerical experiments are important to check the feasibility of the numerical methods and the applicability of the error analysis.

However, since we discretize in space as well, we cannot observe the semi-discrete in time error on the left-hand sides of (4.10) and (4.14) directly in the numerical experiments. Therefore, we measure the fully-discrete error on the left-hand side of (5.24) and choose reasonably fine spatial discretizations that in some sense “come close” to the continuous in space problems. Furthermore, we often study problems, for which the exact solution is known and is at each point in time from the finite element space, so that no discretization error in space occurs.

To somewhat support our claim, that the semi-discrete in time error analysis holds independently of any admissible spatial discretization, we occasionally examine numerical experiments for which we keep all parameters fixed other than the refinement level of the spatial mesh — though for ease of implementation, we do not change the spatial discretization between individual time steps. However, by changing only the level of spatial discretization between different runs of an experiment, we can demonstrate how step size restrictions and discretization error in time are — at least for some problems — largely unaffected by the norm and the condition number of the stiffness matrix. In particular, this suggests that for some problems and below a certain time step size, which depends on the given problem and its exact solution, we might not have to fulfill a CFL-type condition.

In the following sections, we will often speak of W-methods that use *inexact stiffness matrices*. By that we mean, of course, W-methods for which in a fully-discrete algorithm, the exact Fréchet derivatives at previously calculated fully-discrete solutions are replaced with linear operators that only *approximate* those exact Fréchet derivatives. In the fully-discrete algorithm, those approximate operators are then spatially discretized by means of our finite element method. The matrix resulting from that discretization is what we mean by the *inexact stiffness matrix*. In contrast to that, ROW methods always use *exact stiffness matrices* resulting from the spatial discretization of the *exact* Fréchet derivatives.

6.1 A Convection-Diffusion Problem with Known Solution

We begin our numerical experiments with a linear convection-diffusion problem — though in practice, one would, of course, not use W-methods with inexact stiffness matrices for linear problems and fixed spatial discretization, since in that case the stiffness matrix is the same at each time step also for ROW methods. We choose this experiment, however, mainly to isolate the effects that occur, when in a W-method, lower-order convective terms are dropped from the stiffness matrix.

Throughout this section, let $\Omega := (0, 1)^2$, $T \in \mathbb{R}_{>0}$, $\nu \in \mathbb{R}_{>0}$, $b \in L^\infty(\Omega, \mathbb{R})^2$ and $c \in L^\infty(\Omega, \mathbb{R})$. Now we look at the following problem:

Problem 6.1.1. *Let $u_0 \in L^2(\Omega)$ and let $G(t) \in H^{-1}(\Omega)$ for all $t \in (0, T]$. Then we seek a $u \in H^4(0, T; H_0^1((0, 1)^2))$ with*

$$\begin{aligned} \frac{du}{dt}(t) - \nu \Delta u(t) + (b \cdot \nabla)u(t) + cu(t) &= G(t) && \text{for all } t \in (0, T], \\ u(0) &= u_0. \end{aligned}$$

Our goal here is to test the performance of Rosenbrock-type methods and study how the error analysis from previous chapters applies. Because of that, we choose an exact solution that fits into the formulation of problem 6.1.1 above and then calculate the right-hand side G and the initial condition u_0 accordingly. This certainly does not reflect the procedure in actual applications, where the solution is not known. Here, it allows us to measure the errors exactly and it also enables us to choose a solution, which is at each point in time from the finite element space, so that no discretization error in space occurs.

The exact solution we choose is:

$$(0, 1)^2 \times [0, T] \ni (x, y, t) \mapsto u(x, y, t) := x(x-1)y(y-1) \sin(2\pi t). \quad (6.1)$$

We test all methods from section 4.6 by essentially using the semi-discrete (in time) scheme 4.5.1, only that in our fully-discrete algorithms, right-hand sides depend not on the semi-discrete solutions from previous time steps and stages, but on the corresponding fully-discrete ones.

We denote the ROS2 method with $\gamma = 1 + \frac{1}{\sqrt{2}}$ by ROS2p and the ROS2 method with $\gamma = 1 - \frac{1}{\sqrt{2}}$ by ROS2m. For comparative purposes, we also test the Crank-Nicolson method — using the same ordering of temporal and spatial discretization via Rothe's method as with the Rosenbrock-type methods and also using the same spatial discretization. In the tables and figures below we denote the Crank-Nicolson method in short by CN.

We examine all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$. For the spatial discretization, we cover Ω with a mesh of squares that have sides of length 2^{-l} for some $l \in \{4, 5, 6, 7\}$, and then, based on that mesh, we utilize the Q_2 -elements as described in sections 5.1 and 5.1.1.

Furthermore, for implementation purposes we keep the spatial discretization the same at each time step — even though changing spatial discretizations between time steps might be desirable for some applications and could also make it easier to test the semi-discrete error estimates from theorems 4.5.1 and 4.5.2.

Now let $u_{m,h}$ be the so obtained fully discrete numerical solution — with h being the mesh width and $m \in \{0, 1, \dots, M := \frac{T}{\tau}\}$. Also let $t_m := m\tau$ for all $m \in \{0, 1, \dots, M\}$ as usual. In the following figures and tables of this section, we then refer to the error

$$\max_{0 \leq m \leq M} \|u_{m,h} - u(t_m)\|_{L^2}$$

as the $l^\infty L^2$ -error and to the error

$$\left(\tau \sum_{m=1}^M \|u_{m,h} - u(t_m)\|_{H^1}^2 \right)^{\frac{1}{2}}$$

as the $l^2 H^1$ -error. The respective numerical convergence orders for those errors between successive time step sizes are always denoted by q_{num} , with the average numerical convergence order over all time step sizes, for which feasible solutions were computed, being denoted by \bar{q}_{num} .

Even though the spatial discretization is the same at each time step and the above problem 6.1.1 is linear, i.e., the stiffness matrix is also the same at each time step, we want to examine the methods from section 4.6 not just as ROW methods (denoted below with ROW at the end of the name of the method), but also as W-methods with inexact stiffness matrices (denoted below with W at the end of the name of the method).

More specifically, for our W-method variants here we do not build the full stiffness matrix resulting from the Q_2 -discretization of the diffusive term **and** the convective term, but instead we just build the matrix which results from the Q_2 -discretization of **only** the diffusive term, i.e., we omit the convective part. This, of course, does not really save any computing time, as the convective part is the same at each time step, but it does give some insight into what happens, when in W-methods lower order terms of the equation are only handled explicitly rather than implicitly. In addition, we also examined this way of applying W-methods in our theoretical studies — see the first subsection of section 4.5.1 for details.

In the first experiment of this section, we opt for the squares of the spatial mesh to have sides of length 2^{-5} and we choose $T = 1$, $\nu = 1$, $b = (1, 1)^T$, $c = 0$. Here are the resulting $l^\infty L^2$ -errors:

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}
0	9.96e-03	7.10e-03	1.45e-03	2.85e-03	2.35e-04
1	1.25e-03 (2.99)	1.20e-03 (2.57)	3.25e-04 (2.16)	3.98e-04 (2.84)	4.01e-05 (2.55)
2	3.54e-04 (1.82)	2.37e-04 (2.34)	8.30e-05 (1.97)	6.05e-05 (2.72)	7.09e-06 (2.50)
3	2.99e-04 (0.24)	4.99e-05 (2.24)	2.06e-05 (2.01)	9.02e-06 (2.74)	1.12e-06 (2.66)
4	1.36e-04 (1.14)	1.12e-05 (2.16)	5.14e-06 (2.00)	1.28e-06 (2.82)	1.61e-07 (2.79)
5	4.78e-05 (1.51)	2.63e-06 (2.09)	1.28e-06 (2.00)	1.73e-07 (2.89)	2.19e-08 (2.88)
6	1.45e-05 (1.72)	6.37e-07 (2.05)	3.21e-07 (2.00)	2.25e-08 (2.94)	2.85e-09 (2.94)
7	4.01e-06 (1.85)	1.57e-07 (2.02)	8.03e-08 (2.00)	2.88e-09 (2.97)	3.64e-10 (2.97)
8	1.06e-06 (1.92)	3.88e-08 (2.01)	2.01e-08 (2.00)	3.64e-10 (2.98)	4.60e-11 (2.98)
\bar{q}_{num}	1.65	2.19	2.02	2.86	2.79

Table 6.1: A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$
0	2.33e-02	7.45e-03	2.02e-03	3.16e-03	9.03e-04	1.45e-03
1	1.13e-02 (1.04)	2.04e-03 (1.87)	5.45e-04 (1.89)	6.58e-04 (2.26)	2.27e-04 (1.99)	3.25e-04
2	5.14e-03 (1.14)	5.73e-04 (1.83)	1.39e-04 (1.97)	1.79e-04 (1.88)	5.25e-05 (2.11)	8.30e-05
3	2.06e-03 (1.32)	1.52e-04 (1.92)	3.74e-05 (1.89)	5.14e-05 (1.80)	9.86e-06 (2.41)	2.06e-05
4	7.38e-04 (1.48)	3.92e-05 (1.95)	9.49e-06 (1.98)	1.36e-05 (1.92)	1.63e-06 (2.60)	5.14e-06
5	2.35e-04 (1.65)	9.99e-06 (1.97)	2.32e-06 (2.03)	3.34e-06 (2.02)	2.44e-07 (2.74)	1.28e-06
6	6.80e-05 (1.79)	2.53e-06 (1.98)	5.58e-07 (2.06)	7.95e-07 (2.07)	3.46e-08 (2.82)	3.21e-07
7	1.86e-05 (1.87)	6.35e-07 (1.99)	1.34e-07 (2.06)	1.87e-07 (2.09)	4.76e-09 (2.86)	8.03e-08
8	4.89e-06 (1.93)	1.59e-07 (1.99)	3.25e-08 (2.05)	4.42e-08 (2.08)	6.47e-10 (2.88)	2.01e-08
\bar{q}_{num}	1.53	1.94	1.99	2.02	2.55	2.02

Table 6.2: A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

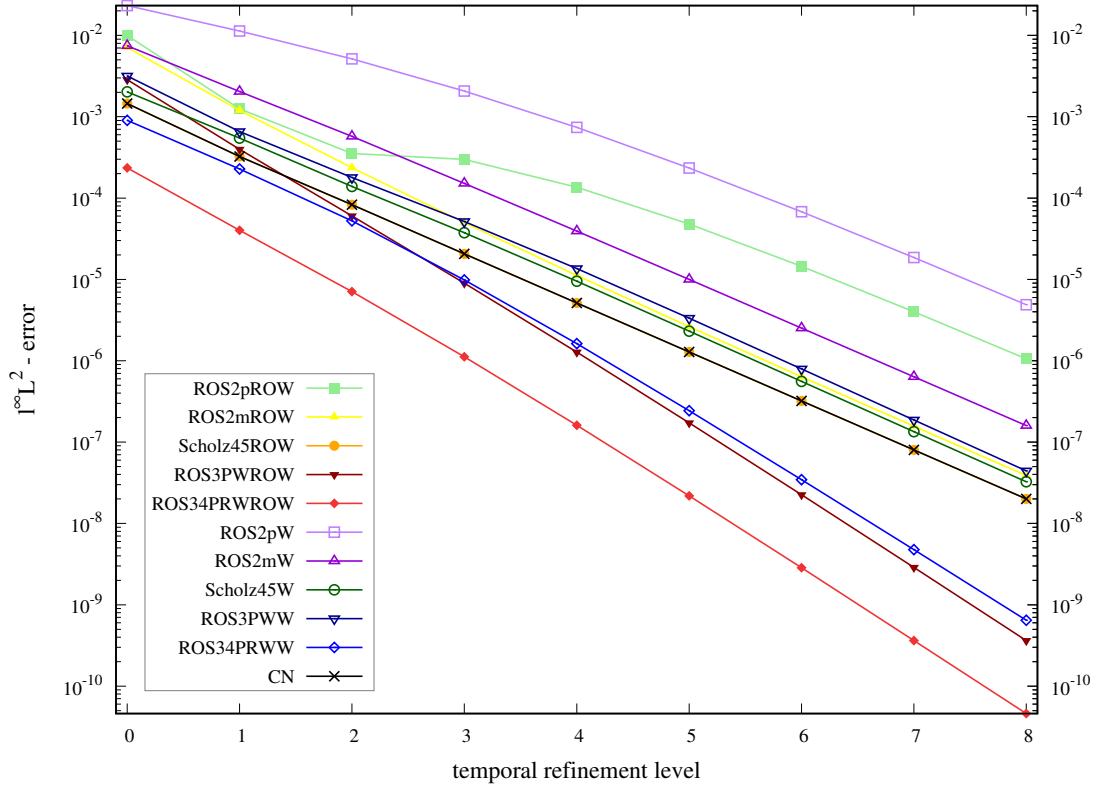


Figure 6.1: A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

We can see that for small time steps, all methods more or less behave as in the classical ODE theory, i.e., their numerical $l^\infty L^2$ -convergence orders are the same as their corresponding classical ODE orders given in theorem 4.3.1 — a theorem which does not account for effects induced by stiffness, such as order reduction and stability problems. Of note here is especially the method ROS3PW. For small time steps it has numerical order 3 as a ROW method but only numerical order 2 as the W-method variant without the convective term in the stiffness matrix. This is in agreement with theorem 4.3.1 and the order conditions stated therein: ROS3PW does fulfill all classical ROW method conditions up to order 3 but does **not** fulfill all classical W-method conditions up to order 3.

For large to medium time steps, the situation is not as clear. Most methods show a dip in numerical $l^\infty L^2$ -convergence order around the time step sizes $\tau = 0.2 \cdot 2^{-2}$ and $\tau = 0.2 \cdot 2^{-3}$. The author's interpretation is that the error bounds from the classical ODE theory and the error bounds from the semi-discrete theory in theorems 4.5.1 and 4.5.2 have a significant influence on the $l^\infty L^2$ -error in different but overlapping regimes of time step sizes — leading to these numerical convergence order phenomena.

From a theoretical standpoint it is unfortunate that our results do not exhibit in any time step size regime a behavior that is clearly identifiable as being dominated by the semi-discrete

error bounds. That this does not happen is understandable, though, as we did not show or claim that the error bounds in theorems 4.5.1 and 4.5.2 are sharp, and we also did not deliberately pick spatial discretizations (possibly changing from time step to time step), methods or problems to test the sharpness of those error bounds. In addition, it is also conceivable that for some methods and for a spatial discretization which is not fine enough, the smallest time step size, which is needed for the semi-discrete error bounds to be valid, is smaller than the time step size, which is needed for the classical error bounds to have a large effect. In that case, the semi-discrete convergence behavior might not be observable at all.

We want to emphasize, though, that for the majority of the time step sizes and almost all methods (the exception being the method ROS2p), the numerical $l^\infty L^2$ -convergence order roughly reaches the order of convergence given in the semi-discrete theory — this is best seen in the average numerical $l^\infty L^2$ -convergence orders \bar{q}_{num} at the bottom of the above tables.

A positive exception is the method ROS34PRWW (ROS34PRW as W-method, i.e., without the convective part in the stiffness matrix). In the classical ODE theory, ROS34PRW has order 3 when used as W-method with approximate Jacobian. Theorem 4.5.2, on the other hand, only guarantees order 2 for W-methods and small enough time steps in the semi-discrete case. And in the experiments we observe a numerical $l^\infty L^2$ -convergence order of almost 3 for most time step sizes. This is a first hint, that the error bound from theorem 4.5.2 might not be sharp. When examining the $l^2 H^1$ -errors below, we will see a more clear indication of that.

As already mentioned, the method ROS2p does not quite show a numerical $l^\infty L^2$ -convergence order that matches the convergence order given in the semi-discrete theory. For larger time steps, that is also the case for the method ROS34PRW. This most likely means, that the minimum time step size needed for the semi-discrete error estimates to be valid is a bit smaller for these methods than for the other methods.

Looking at the absolute errors, we can say that the ROW methods (full stiffness matrix) perform *significantly*, though not *much* better than our W-methods without the convective part in the stiffness matrix — with the exception of ROS3PW, which, as we already mentioned above, loses an order when the convective term is dropped from the stiffness matrix. However, the method ROS34PRWW (ROS34PRW as W-method without the convective part in the matrix) does still display a very good performance, almost being on par with the method ROS3PWROW (ROS3PW as ROW method).

Further examining the results, we observe that Scholz45 is our best order 2 method, both as ROW method and as W-method, i.e., Scholz45ROW is the best order 2 ROW method and Scholz45W is the best of our order 2 W-method variants that handle the convective term only explicitly. Note that Scholz45ROW produces the exact same errors as the Crank-Nicolson method. In fact, it is easy to see that for linear problems, such as the current one, Crank-Nicolson and Scholz45ROW are indeed equivalent.

Among the two order 3 methods, ROS34PRW (a four stage method) clearly shows the better performance than the method ROS3PW (a three stage method) — both as ROW method and as W-method.

Next, we look at the $l^2 H^1$ -errors. With a few exceptions, those errors do not display a significantly different behavior than the $l^\infty L^2$ -errors. Hence, we only expand on the above interpretation

and explanation of the $l^\infty L^2$ -errors when new findings demand it.

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}
0	3.27e-02	2.28e-02	5.37e-03	9.50e-03	7.39e-04
1	4.31e-03 (2.92)	4.05e-03 (2.49)	1.08e-03 (2.32)	1.26e-03 (2.92)	1.29e-04 (2.52)
2	1.21e-03 (1.83)	7.73e-04 (2.39)	2.51e-04 (2.10)	1.84e-04 (2.77)	2.18e-05 (2.57)
3	9.53e-04 (0.35)	1.60e-04 (2.27)	6.13e-05 (2.03)	2.69e-05 (2.77)	3.38e-06 (2.69)
4	4.30e-04 (1.15)	3.55e-05 (2.17)	1.52e-05 (2.01)	3.78e-06 (2.83)	4.85e-07 (2.80)
5	1.51e-04 (1.51)	8.29e-06 (2.10)	3.78e-06 (2.01)	5.08e-07 (2.89)	6.55e-08 (2.89)
6	4.57e-05 (1.72)	2.00e-06 (2.05)	9.44e-07 (2.00)	6.62e-08 (2.94)	8.54e-09 (2.94)
7	1.27e-05 (1.85)	4.90e-07 (2.03)	2.36e-07 (2.00)	8.47e-09 (2.97)	1.09e-09 (2.97)
8	3.34e-06 (1.92)	1.21e-07 (2.01)	5.89e-08 (2.00)	1.07e-09 (2.98)	1.39e-10 (2.98)
\bar{q}_{num}	1.66	2.19	2.06	2.89	2.79

Table 6.3: A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$
0	7.11e-02	2.85e-02	9.79e-03	1.27e-02	4.98e-03	5.37e-03
1	3.46e-02 (1.04)	7.70e-03 (1.89)	2.47e-03 (1.99)	3.08e-03 (2.05)	1.32e-03 (1.91)	1.08e-03
2	1.54e-02 (1.17)	2.10e-03 (1.87)	6.93e-04 (1.83)	9.31e-04 (1.72)	3.05e-04 (2.12)	2.51e-04
3	6.23e-03 (1.31)	5.77e-04 (1.87)	1.89e-04 (1.88)	2.71e-04 (1.78)	6.27e-05 (2.28)	6.13e-05
4	2.24e-03 (1.47)	1.59e-04 (1.86)	4.87e-05 (1.95)	7.22e-05 (1.91)	1.20e-05 (2.39)	1.52e-05
5	7.28e-04 (1.62)	4.39e-05 (1.86)	1.21e-05 (2.01)	1.81e-05 (2.00)	2.22e-06 (2.43)	3.78e-06
6	2.19e-04 (1.73)	1.21e-05 (1.86)	2.94e-06 (2.04)	4.35e-06 (2.05)	4.09e-07 (2.44)	9.44e-07
7	6.32e-05 (1.79)	3.30e-06 (1.87)	7.08e-07 (2.05)	1.03e-06 (2.08)	7.51e-08 (2.45)	2.36e-07
8	1.79e-05 (1.82)	8.88e-07 (1.90)	1.71e-07 (2.05)	2.44e-07 (2.08)	1.37e-08 (2.46)	5.89e-08
\bar{q}_{num}	1.49	1.87	1.98	1.96	2.31	2.06

Table 6.4: A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

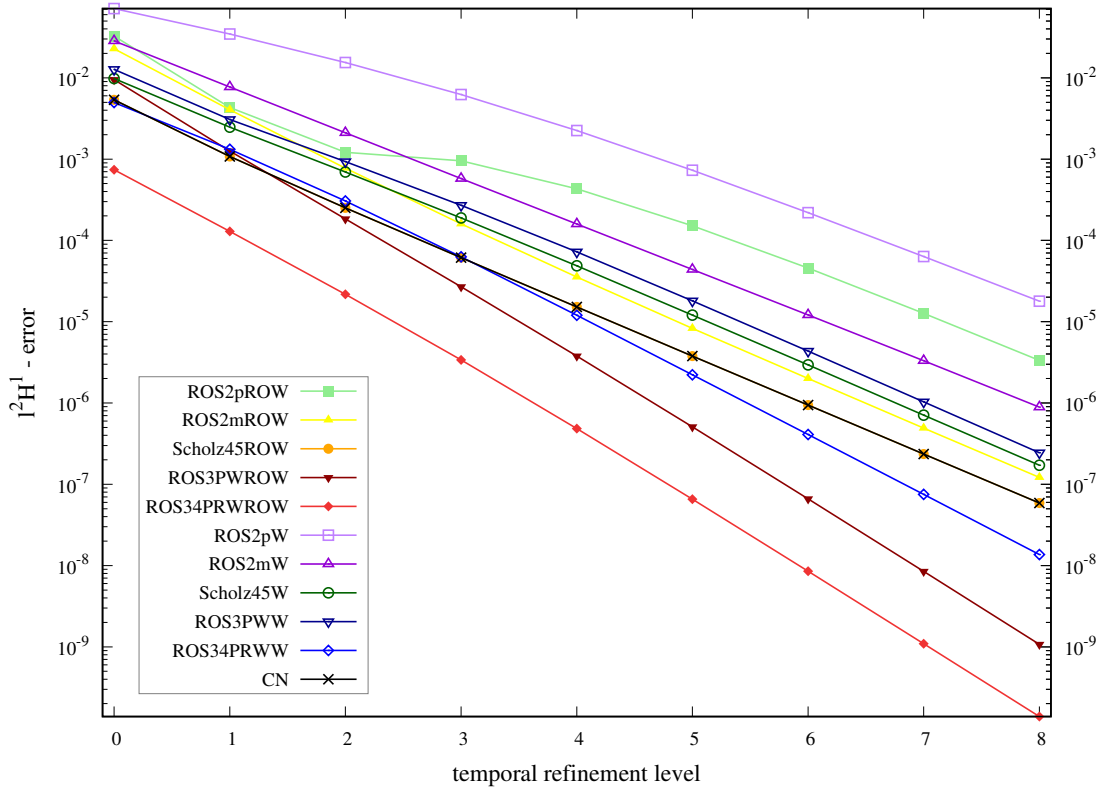


Figure 6.2: A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

As already mentioned, the $l^2 H^1$ -errors show largely the same behavior as the $l^\infty L^2$ -errors (see the discussion below figure 6.1 for a somewhat detailed review of the $l^\infty L^2$ -error results).

A very notable exception to that can be observed for the method ROS34PRWW (ROS34PRW as W-method without the convective term in the matrix), which for medium and small time steps reached a numerical $l^\infty L^2$ -convergence order of almost 3. Its numerical $l^2 H^1$ -convergence order, on the other hand, does not exceed 2.46 — even for the smallest time steps tested. A possible explanation is, that for ROS34PRWW the minimum time step size required for the classical error bounds to have a significant influence on the $l^2 H^1$ -error is very small in this case compared to the minimum time step size required for any semi-discrete error bounds to have an influence. This would also mean, though, that the semi-discrete convergence order of 2, which is provided by theorem 4.5.2, might not be sharp for this problem and method — as we observe a fractional order of convergence larger than 2 but smaller than 3.

Fractional temporal orders of convergence for time-stepping schemes applied to parabolic PDEs have been theoretically predicted and also numerically observed — for the theoretical analysis of that phenomenon see for example the paper [45] by Ostermann and Roche or the papers [40, 42, 41] by Lubich and Ostermann. Theorem 4.5.2 was in fact obtained from the paper [41]. However, in theorem 4.5.2 and the paper [41], the W-methods are only required to have at least

classical order 2 — and ROS34PRW has classical order 3. Formulating sharp semi-discrete error bounds for some W-methods of higher order could thus be a task for the future.

Now we slightly change the experiment in that we look at a very similar problem with the same exact solution and the same spatial and temporal discretization as before, with the only difference in the problem formulation being that we now look at a convection-dominated version, i.e., we change the diffusion coefficient from $\nu = 1$ to the considerably smaller size of $\nu = 10^{-4}$. Notice, though, that our exact solution is not affected by that, as it is independent of ν . We first look at the resulting $l^\infty L^2$ -errors:

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}
0	1.64e-02	4.58e-03	6.34e-03	4.22e-03	1.88e-03
1	9.37e-03 (0.80)	1.27e-03 (1.84)	1.72e-03 (1.88)	7.80e-04 (2.43)	2.94e-04 (2.68)
2	4.48e-03 (1.07)	3.22e-04 (1.99)	4.45e-04 (1.96)	1.12e-04 (2.80)	3.95e-05 (2.90)
3	1.48e-03 (1.60)	8.18e-05 (1.98)	1.12e-04 (2.00)	1.47e-05 (2.93)	5.06e-06 (2.97)
4	4.08e-04 (1.86)	2.05e-05 (1.99)	2.79e-05 (2.00)	1.86e-06 (2.98)	6.37e-07 (2.99)
5	1.06e-04 (1.95)	5.14e-06 (2.00)	6.98e-06 (2.00)	2.34e-07 (2.99)	7.99e-08 (3.00)
6	2.67e-05 (1.98)	1.29e-06 (2.00)	1.75e-06 (2.00)	2.93e-08 (3.00)	1.00e-08 (3.00)
7	6.71e-06 (1.99)	3.22e-07 (2.00)	4.36e-07 (2.00)	3.66e-09 (3.00)	1.25e-09 (3.00)
8	1.68e-06 (2.00)	8.05e-08 (2.00)	1.09e-07 (2.00)	4.58e-10 (3.00)	1.56e-10 (3.00)
\bar{q}_{num}	1.66	1.97	1.98	2.89	2.94

Table 6.5: A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$
0	1.94e+03	8.55e+03	6.56e+03	3.48e+03	1.26e+11	6.34e-03
1	1.08e+08	1.71e+09	1.07e+09	5.45e+08	3.23e+15	1.73e-03
2	1.43e+14	1.88e+15	1.12e+15	9.58e+14	2.13e+15	4.45e-04
3	9.16e+14	4.56e+14	1.54e+15	8.44e+14	2.99e+05	1.12e-04
4	1.35e+05	1.43e+05	1.38e+05	1.92e+05	3.70e-07	2.79e-05
5	1.40e-05	1.40e-05	1.40e-05	1.30e-05	4.61e-08 (3.00)	6.98e-06
6	3.50e-06 (2.00)	3.50e-06 (2.00)	3.25e-06 (2.00)	4.47e-06 (2.00)	5.76e-09 (3.00)	1.74e-06
7	8.74e-07 (2.00)	8.75e-07 (2.00)	8.12e-07 (2.00)	1.12e-06 (2.00)	7.20e-10 (3.00)	4.36e-07
8	2.19e-07 (2.00)	2.19e-07 (2.00)	2.19e-07 (2.00)	2.03e-07 (2.00)	9.00e-11 (3.00)	1.09e-07
\bar{q}_{num}	2.00	2.00	2.00	2.00	3.00	1.98

Table 6.6: A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

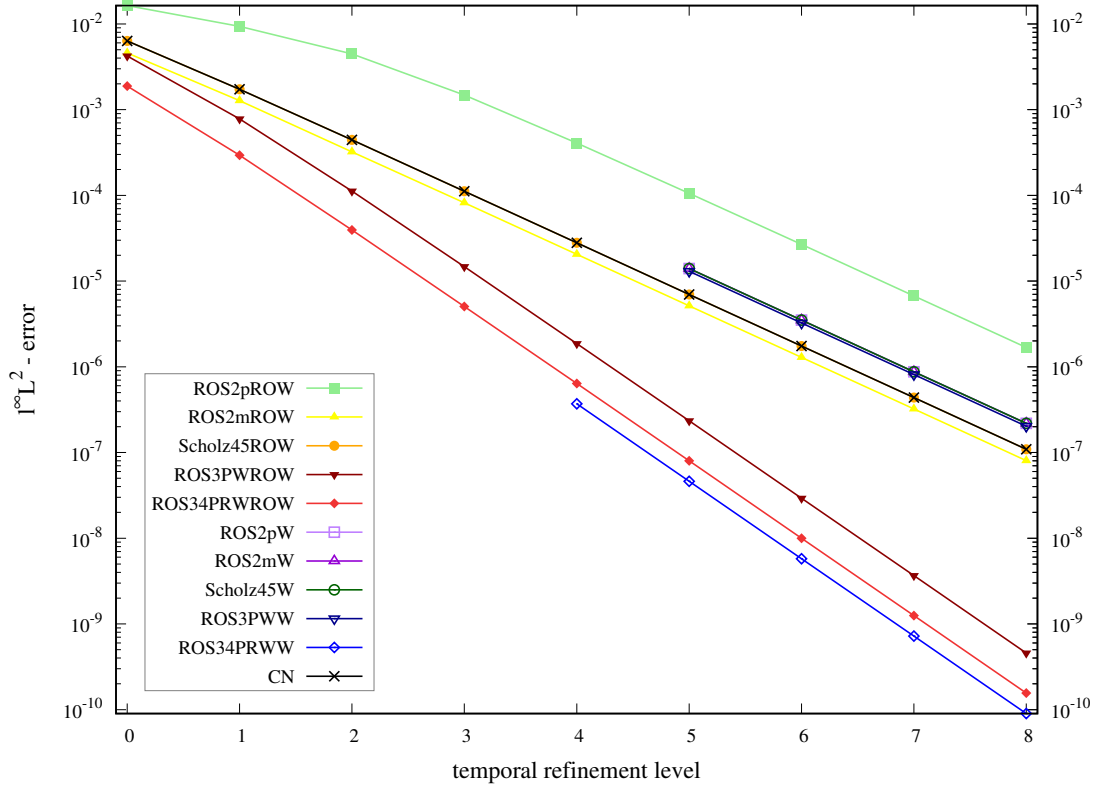


Figure 6.3: A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

When looking at these $l^\infty L^2$ -errors for our convection-dominated problem, we can see that the ROW methods behave very similarly as for the convection-diffusion problem that is not convection-dominated (see the discussion below figure 6.1 for a somewhat detailed review of those results).

A notable difference for the ROW methods in the convection-dominated case ($\nu = 10^{-4}$) compared to the standard case ($\nu = 1$) seems to be that the numerical $l^\infty L^2$ -convergence orders in the convection-dominated case are a bit larger for small time steps, but smaller for very large time steps. The absolute $l^\infty L^2$ -errors in the convection-dominated case are overall a bit larger as well for ROW methods and also for the Crank-Nicolson method.

For our W-methods, however, we observe a very remarkable difference in the convection-dominated case, compared to the standard case. Omitting the convective term from the stiffness matrix when we have $\nu = 10^{-4}$ seems to generate a stability problem, i.e., there is a minimum time step size, before which these W-methods do not work at all. The methods ROS2pW, ROS2mW, Scholz45W and ROS3PWW produce their first feasible solution for the time step size $\tau = 0.2 \cdot 2^{-5}$. The method ROS34PRWW seems to be a bit more stable, i.e., $\tau = 0.2 \cdot 2^{-4}$ already works.

For the time step sizes that do yield viable solutions, the numerical $l^\infty L^2$ -convergence orders

are for all methods exactly the same as the corresponding classical ODE orders given in theorem 4.3.1. It is also noteworthy, that the absolute $l^\infty L^2$ -errors in those small time steps with viable solutions are for the methods ROS2mW, Scholz45W and ROS3PWW a bit *larger* and for the methods ROS2pW and ROS34PRWW a bit *smaller* in the convection-dominated-case than in the standard case.

Another remarkable finding is that the $l^\infty L^2$ -errors at those time steps with viable solutions are *identical* for the methods ROS2pW, ROS2mW and Scholz45W — with the method ROS3PWW also producing almost the same error.

Judging by the $l^\infty L^2$ -errors in the convection-dominated case, it seems clear that the method ROS34PRWW performs best among the tested W-methods, both in terms of stability and accuracy.

Below, we present the $l^2 H^1$ -errors for our convection-dominated convection-diffusion problem.

	ROS2pROW		ROS2mROW		Scholz45ROW		ROS3PWWROW		ROS34PRWROW	
k	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}
0	5.88e−01		1.24e−01		2.51e−01		1.73e−01		1.12e−01	
1	4.16e−01	(0.50)	3.50e−02	(1.82)	7.28e−02	(1.78)	3.58e−02	(2.27)	1.60e−02	(2.81)
2	2.15e−01	(0.95)	8.90e−03	(1.98)	1.84e−02	(1.98)	5.50e−03	(2.70)	2.05e−03	(2.96)
3	7.06e−02	(1.61)	2.22e−03	(2.01)	4.58e−03	(2.01)	7.32e−04	(2.91)	2.55e−04	(3.01)
4	1.89e−02	(1.90)	5.52e−04	(2.01)	1.14e−03	(2.01)	9.27e−05	(2.98)	3.17e−05	(3.01)
5	4.74e−03	(2.00)	1.37e−04	(2.00)	2.83e−04	(2.01)	1.16e−05	(3.00)	3.94e−06	(3.01)
6	1.18e−03	(2.01)	3.43e−05	(2.00)	7.05e−05	(2.00)	1.45e−06	(3.00)	4.91e−07	(3.00)
7	2.92e−04	(2.01)	8.57e−06	(2.00)	1.76e−05	(2.00)	1.81e−07	(3.00)	6.13e−08	(3.00)
8	7.27e−05	(2.01)	2.14e−06	(2.00)	4.40e−06	(2.00)	2.27e−07	(3.00)	7.65e−09	(3.00)
\bar{q}_{num}	1.62		1.98		1.97		2.86		2.98	

Table 6.7: A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW		ROS2mW		Scholz45W		ROS3PWW		ROS34PRWW		CN
k	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$
0	7.48e+04		3.86e+05		2.87e+05		1.52e+05		6.71e+12		2.51e-01
1	3.90e+09		6.65e+10		4.11e+10		2.09e+10		1.76e+17		7.28e-02
2	3.81e+15		5.25e+16		3.10e+16		2.63e+16		1.41e+17		1.84e-02
3	3.07e+16		1.95e+16		5.91e+16		3.25e+16		6.28e+06		4.58e-03
4	2.39e+06		2.43e+06		2.36e+06		3.30e+06		1.48e-05		1.14e-03
5	5.61e-04		5.61e-04		5.61e-04		5.61e-04		1.84e-06 (3.01)		2.83e-04
6	1.40e-04 (2.00)		1.40e-04 (2.00)		1.40e-04 (2.00)		2.38e-04 (2.00)		2.29e-07 (3.00)		7.05e-05
7	3.51e-05 (2.00)		3.51e-05 (2.00)		3.51e-05 (2.00)		3.51e-05 (2.00)		2.86e-08 (3.00)		1.76e-05
8	8.78e-06 (2.00)		8.78e-06 (2.00)		8.78e-06 (2.00)		8.78e-06 (2.00)		3.58e-09 (3.00)		4.40e-06
\bar{q}_{num}	2.00		2.00		2.00		2.00		3.00		1.97

Table 6.8: A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

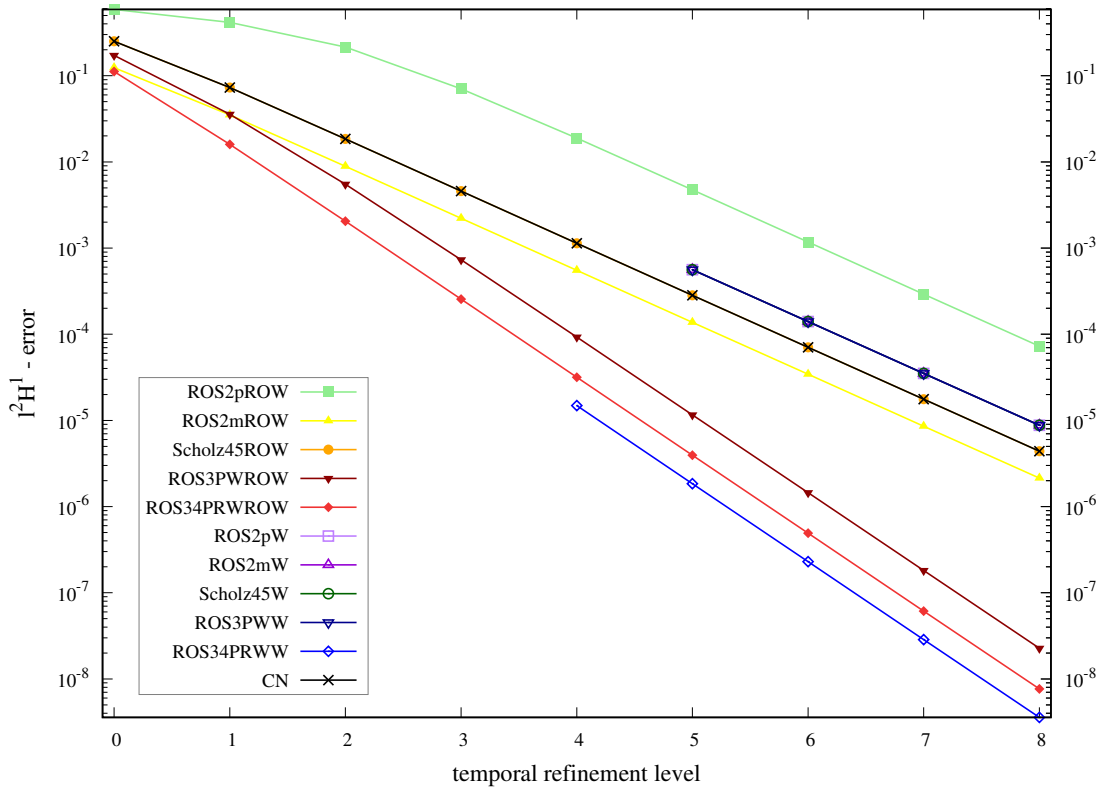


Figure 6.4: A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

We can see that in the convection-dominated case, these l^2H^1 -errors do not add much insight to what we gathered above from the $l^\infty L^2$ -errors. Naturally, we do not repeat those arguments here.

One thing, which is noteworthy is that in the standard case ($\nu = 1$), the method ROS34PRWW (ROS34PRW as W-method without convective term in the stiffness matrix) showed even for small time steps only a fractional numerical l^2H^1 -convergence order of around 2.46 — in the convection-dominated case ($\nu = 10^{-4}$), on the other hand, the numerical l^2H^1 -convergence order between time steps for which viable solutions are produced is exactly 3 for the method ROS34PRWW.

We just discovered that for our convection-dominated convection-diffusion problems, the W-methods without convective term in the stiffness matrix do not work when large time steps are used. Now we want to briefly test, whether the minimum time step size, from which on those methods do produce viable solutions, depends on the fineness of the spatial discretization, i.e., we want to see, whether we have to fulfill a CFL-type condition here.

To do that, we employ numerical experiments, very similar to the ones above. We still look at problem 6.1.1, and we still set $b = (1, 1)^T$, $c = 0$. We also use the same spatial and temporal discretization techniques that we described at the beginning of this section.

However, for the subsequent experiments, we do not keep the spatial discretization fixed to always having squares with sides of length 2^{-5} as we did above. Instead, we vary the spatial refinement level k between 4, 5, 6, 7 and 8 — with the corresponding side length of the squares, which we use for the Q_2 -elements, being 2^{-k} . In addition, we do not use a maximum length of $T = 1$ for the time interval, but rather lengthen it to $T = 10$ to improve our ability of detecting stability issues in our time-stepping schemes. We do still use the same exact solution (6.1) and set the right-hand side accordingly.

Now we choose the diffusion coefficient ν *just* small enough, so that the tested time-stepping method fails to produce a viable solution for the spatial refinement level 4 (i.e., the squares of the spatial mesh have sides of length 2^{-4}) and the temporal refinement level 0 (i.e., the time step size is $\tau = 0.2 \cdot 2^{-0}$).

Then we increase the temporal refinement level twice, i.e., we decrease the time step size, to see if we obtain viable solutions. After that, we keep the temporal refinement level *fixed* at $k = 2$ but increase the spatial refinement level to see if that leads to stability problems. We test Scholz45W and ROS34PRWW, two of our best performing methods. The results are as follows:

l	k	Scholz45W with $\nu = 0.05$		ROS34PRWW with $\nu = 0.025$	
		$l^\infty L^2$ - error	$l^2 H^1$ - error	$l^\infty L^2$ - error	$l^2 H^1$ - error
4	0	3.1916e+10	3.1047e+11	3.0261e+07	4.2111e+08
4	1	4.8405e+00	1.1772e+02	2.7048e+00	6.9035e+01
4	2	4.2144e−04	8.3618e−03	3.9889e−05	1.5324e−03
5	2	4.2024e−04	8.3603e−03	4.0300e−05	1.6197e−03
6	2	4.2010e−04	8.3595e−03	4.0023e−05	1.6395e−03
7	2	4.2009e−04	8.3594e−03	3.9926e−05	1.6420e−03

Table 6.9: A convection-diffusion problem with known solution. Testing independence on spatial refinement for Scholz45W and ROS34PRWW: l denotes the spatial refinement level (with the squares of the spatial mesh having sides of length 2^{-l}) and k denotes the temporal refinement level (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

As we can see, even though the problem is chosen in such a way that the experiments are conducted near a maximally admissible time step size, substantially increasing the fineness of the spatial mesh does not require us to further decrease the time step size in order to circumvent stability issues. Furthermore, the $l^\infty L^2$ -errors and the $l^2 H^1$ -errors are not significantly influenced by the fineness of the spatial mesh either.

6.2 A Reaction-Diffusion Problem with Known Solution

In the section above, we looked at a linear problem — mainly to isolate the effects that occur, when in a W-method, lower-order convective terms are dropped from the stiffness matrix.

In applications with linear problems and fixed spatial discretization, however, the stiffness matrix is the same at each time step. Hence, in that case W-methods do not promise to have any advantage over ROW methods because those would also require a stiffness matrix to be built only once during the whole time-stepping procedure. The situation changes when we introduce a nonlinearity into the equation. In ROW methods, even with a fixed spatial discretization, one needs to assemble a new stiffness matrix at each time step — for very fine spatial discretizations that can become very computationally costly.

In this section, we thus look at an example, in which we try to isolate the effects of having a nonlinearity in the equation. This also means that we drop any convective terms. In the following section, where we study various numerical experiments on the Navier-Stokes equations, we then have a situation where both influences occur combined: lower order *convective* terms that are also *nonlinear*.

Throughout this section, let $\Omega := (0, 1)^2$, $T \in \mathbb{R}_{>0}$ and $\nu \in \mathbb{R}_{>0}$. Now we look at the following problem:

Problem 6.2.1. Let $u_0 \in L^2(\Omega)$ and let $G(t) \in H^{-1}(\Omega)$ for all $t \in (0, T]$.

Then we seek a $u \in H^4(0, T; H_0^1((0, 1)^2))$ with

$$\begin{aligned} \frac{du}{dt}(t) - \nu \Delta u(t) + (u(t))^2(u(t) - 1) &= G(t) && \text{for all } t \in (0, T], \\ u(0) &= u_0. \end{aligned}$$

As in the experiments on the convection-diffusion problem, our goal here is to test the performance of Rosenbrock-type methods and examine how the error analysis from previous chapters applies. We do this by choosing an exact solution that fits into the formulation of problem 6.2.1 above and then calculating the right-hand side G and the initial condition u_0 accordingly. In practice, the solution is, of course, not known beforehand. Here, *choosing* the exact solution allows us to measure the errors exactly and it also enables us to pick a solution, which is at each point in time from the finite element space, so that no discretization error in space occurs.

The exact solution we choose is the same as the one we chose for the convection-diffusion problem:

$$(0, 1)^2 \times [0, T] \ni (x, y, t) \mapsto u(x, y, t) := x(x - 1)y(y - 1) \sin(2\pi t). \quad (6.2)$$

We again test all methods from section 4.6 by essentially using the semi-discrete (in time) scheme 4.5.1, only that in our fully-discrete algorithms, operators and right-hand sides depend not on the semi-discrete solutions from previous time steps and stages, but on the corresponding fully-discrete ones.

We denote the ROS2 method with $\gamma = 1 + \frac{1}{\sqrt{2}}$ by ROS2p and the ROS2 method with $\gamma = 1 - \frac{1}{\sqrt{2}}$ by ROS2m. And once more, for comparative purposes, we also test the Crank-Nicolson method — using the same ordering of temporal and spatial discretization via Rothe’s method as with the Rosenbrock-type methods and also using the same spatial discretization. We solve any nonlinear systems, which arise when using the Crank-Nicolson method here, with Newton’s method. In the tables and figures below we denote the Crank-Nicolson method in short by CN.

We also use the same spatial and temporal discretization techniques as in the previous experiments on the convection-diffusion problem. This means that we examine all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$ and for the spatial discretization, we cover Ω with a mesh of squares that have sides of length 2^{-l} for some $l \in \{4, 5, 6, 7\}$, and then, based on that mesh, we utilize the Q_2 -elements as described in sections 5.1 and 5.1.1.

For implementation purposes we again keep the spatial discretization the same at each time step — even though changing spatial discretizations between time steps might be desirable for some applications and could also make it easier to test the semi-discrete error estimates from theorems 4.5.1 and 4.5.2.

Now let $u_{m,h}$ be the so obtained fully discrete numerical solution — with h being the mesh width and $m \in \{0, 1, \dots, M := \frac{T}{\tau}\}$. Also let $t_m := m\tau$ for all $m \in \{0, 1, \dots, M\}$ as usual. We then use the same notation as in the previous experiments on the convection-diffusion problem, i.e., in the following figures and tables of this section, we then refer to the error

$$\max_{0 \leq m \leq M} \|u_{m,h} - u(t_m)\|_{L^2}$$

$l^\infty L^2$ -error and to the error

$$\left(\tau \sum_{m=1}^M \|u_{m,h} - u(t_m)\|_{H^1}^2 \right)^{\frac{1}{2}}$$

as the $l^2 H^1$ -error. The respective numerical convergence orders for those errors between successive time step sizes are always denoted by q_{num} , with the average numerical convergence order over all time step sizes, for which feasible solutions were computed, being denoted by \bar{q}_{num} .

For the ROW methods (denoted below with ROW at the end of the name of the method), we have to build at each time step a new stiffness matrix: the Q_2 -discretization of the diffusive term plus the Q_2 -discretization of the Fréchet derivative (at the numerical solution from the last time step) of the nonlinearity.

In this section we also test W-methods (denoted below with W at the end of the name of the method) for which we choose to utilize the *same* approximate stiffness matrix at each time step: the stiffness matrix, which is used by the ROW methods at the *first* time step. With our specific exact solution and corresponding initial condition, this just means that the approximate stiffness matrix used for those W-methods at each time step is merely the Q_2 -discretization of the diffusive term.

In the all experiments of this section, we choose $T = 1$ and we opt for the squares of the spatial mesh to have sides of length 2^{-5} . There are two main sets of experiments, one with diffusion coefficient $\nu = 1$ and one with $\nu = 10^{-4}$. We first look at the results for $\nu = 1$. Here are the $l^\infty L^2$ -errors:

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}
0	9.84e−03	7.15e−03	1.50e−03	2.88e−03	2.23e−04
1	1.21e−03 (3.03)	1.20e−03 (2.57)	3.30e−04 (2.18)	3.99e−04 (2.85)	4.13e−05 (2.43)
2	3.59e−04 (1.75)	2.37e−04 (2.34)	8.39e−05 (1.98)	6.07e−05 (2.72)	7.12e−06 (2.54)
3	3.02e−04 (0.25)	5.00e−05 (2.25)	2.08e−05 (2.01)	9.03e−06 (2.74)	1.12e−06 (2.67)
4	1.37e−04 (1.15)	1.12e−05 (2.16)	5.19e−06 (2.00)	1.28e−06 (2.82)	1.61e−07 (2.80)
5	4.79e−05 (1.51)	2.64e−06 (2.09)	1.30e−06 (2.00)	1.73e−07 (2.89)	2.18e−08 (2.88)
6	1.45e−05 (1.73)	6.38e−07 (2.05)	3.24e−07 (2.00)	2.25e−08 (2.94)	2.84e−09 (2.94)
7	4.00e−06 (1.85)	1.57e−07 (2.02)	8.11e−08 (2.00)	2.87e−09 (2.97)	3.63e−10 (2.97)
8	1.06e−06 (1.92)	3.89e−08 (2.01)	2.03e−08 (2.00)	3.63e−10 (2.98)	4.59e−11 (2.98)
\bar{q}_{num}	1.65	2.19	2.02	2.86	2.78

Table 6.10: A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW		ROS2mW		Scholz45W		ROS3PWW		ROS34PRWW		CN
k	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$
0	2.31e-02		7.46e-03		1.47e-03		2.81e-03		2.51e-04		1.46e-03
1	1.14e-02 (1.03)		2.06e-03 (1.86)		3.35e-04 (2.14)		4.07e-04 (2.79)		4.14e-05 (2.60)		3.28e-04
2	5.19e-03 (1.13)		5.70e-04 (1.85)		8.30e-05 (2.01)		6.22e-05 (2.71)		7.26e-06 (2.51)		8.38e-05
3	2.09e-03 (1.31)		1.50e-04 (1.93)		2.08e-05 (2.00)		9.58e-06 (2.70)		1.11e-06 (2.70)		2.08e-05
4	7.46e-04 (1.48)		3.87e-05 (1.95)		5.20e-06 (2.00)		1.51e-06 (2.67)		1.62e-07 (2.78)		5.19e-06
5	2.37e-04 (1.65)		9.85e-06 (1.97)		1.30e-06 (2.00)		2.56e-07 (2.56)		2.23e-08 (2.86)		1.30e-06
6	6.88e-05 (1.79)		2.49e-06 (1.98)		3.26e-07 (2.00)		4.86e-08 (2.40)		2.93e-09 (2.93)		3.24e-07
7	1.88e-05 (1.87)		6.26e-07 (1.99)		8.16e-08 (2.00)		1.02e-08 (2.25)		3.76e-10 (2.96)		8.11e-08
8	4.93e-06 (1.93)		1.57e-07 (1.99)		2.04e-08 (2.00)		2.32e-09 (2.14)		4.76e-11 (2.98)		2.03e-08
\bar{q}_{num}	1.52		1.94		2.02		2.53		2.79		2.02

Table 6.11: A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

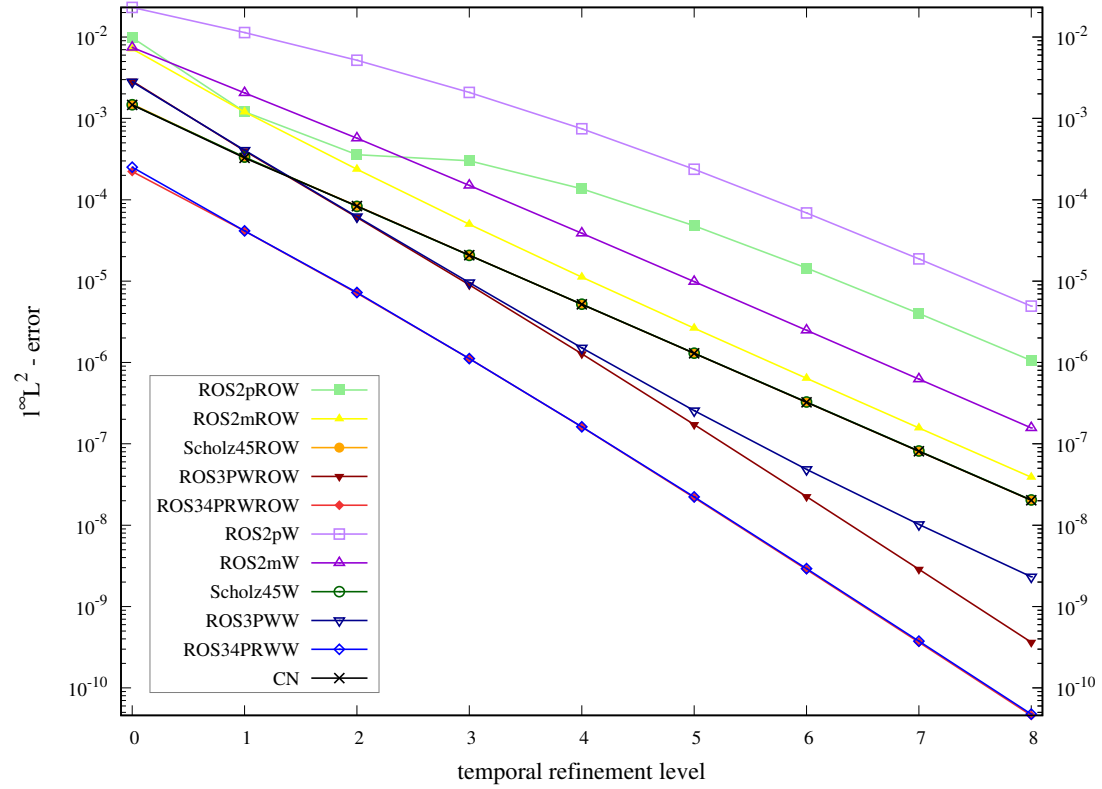


Figure 6.5: A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

Regarding the $l^\infty L^2$ -errors obtained from the reaction-diffusion experiments, we can make similar observations as for the convection-diffusion experiments: for small time steps, all methods more or less behave as in the classical ODE theory, i.e., their numerical $l^\infty L^2$ -convergence orders are the same as their corresponding classical ODE orders given in theorem 4.3.1. The method ROS3PW, for example, has for small time steps numerical order 3 when used as a ROW method, but its numerical $l^\infty L^2$ -convergence order falls to 2 for small time steps when it is used as a W-method with inexact stiffness matrices. This is in agreement with theorem 4.3.1 and the order conditions stated therein: ROS3PW does fulfill all classical ROW method conditions up to order 3 but does **not** fulfill all classical W-method conditions up to order 3.

For large to medium time steps, on the other hand, the numerical $l^\infty L^2$ -convergence order of many methods is significantly worse than their classical order — with the exception being the methods ROS2m and Scholz45, which (both as ROW methods and as W-methods) have classical order 2 and for pretty much all time step sizes also numerical $l^\infty L^2$ -convergence order 2.

The only method, however, that has an *average* numerical $l^\infty L^2$ -convergence order far below the convergence order given in the semi-discrete theory is the method ROS2p (both as ROW method and as W-method), though for small time steps, its numerical $l^\infty L^2$ -convergence order does near 2. The author's interpretation is that for ROS2p, the minimum time step size required for the semi-discrete error bounds to have a significant influence on the numerical order is very small for this method compared to the other methods.

Judging by the absolute $l^\infty L^2$ -errors, it once again seems like Scholz45ROW/W (which both produce almost the exact same $l^\infty L^2$ -errors as the Crank-Nicolson method) are the best order 2 ROW methods/W-methods, respectively, and ROS34PRWROW/W are better than ROS3PWROW/W, respectively — remember, though, that ROS34PRW has four stages, whereas ROS3PW only has three.

Moreover, we find it noteworthy to mention that when comparing for each method the ROW method version with its respective W-method counterpart, one can see that for the methods ROS2p, ROS2m and ROS3PW the ROW method version performs better, whereas for the methods Scholz45 and ROS34PRW, the ROW method and W-method versions produce approximately the same absolute $l^\infty L^2$ -errors.

We find this remarkable because it means that for the methods Scholz45 and ROS34PRW, omitting the contribution from the nonlinearity in our reaction-diffusion example from the stiffness matrix does not have a significant influence on the $l^\infty L^2$ -errors. In practice, especially when using a very fine spatial discretization, i.e., a large stiffness matrix, one might thus save a lot of computing time by only building the stiffness matrix once, instead of at every time step. Additional experiments, involving careful measuring of CPU time, could give a clearer picture of the benefit that W-methods with inexact stiffness matrices provide for reaction-diffusion problems, such as the one we are currently examining.

Next, we show the $l^2 H^1$ -errors for the reaction-diffusion example. Those errors do not display a significantly different behavior than the $l^\infty L^2$ -errors. Hence, we will not expand on the above interpretation and explanation of the $l^\infty L^2$ -errors. The only noteworthy difference is, that for all tested methods (including the Crank-Nicolson method), the absolute $l^2 H^1$ -errors are considerably larger than the absolute $l^\infty L^2$ -errors.

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}
0	3.25e-02	2.26e-02	5.40e-03	9.49e-03	7.49e-04
1	4.08e-03 (2.99)	4.03e-03 (2.49)	1.08e-03 (2.32)	1.26e-03 (2.92)	1.30e-04 (2.53)
2	1.16e-03 (1.82)	7.70e-04 (2.39)	2.52e-04 (2.10)	1.84e-04 (2.78)	2.18e-05 (2.57)
3	9.48e-04 (0.29)	1.60e-04 (2.27)	6.16e-05 (2.04)	2.68e-05 (2.78)	3.38e-06 (2.69)
4	4.29e-04 (1.14)	3.55e-05 (2.17)	1.53e-05 (2.01)	3.77e-06 (2.83)	4.83e-07 (2.80)
5	1.51e-04 (1.51)	8.30e-06 (2.10)	3.80e-06 (2.01)	5.06e-07 (2.89)	6.52e-08 (2.89)
6	4.56e-05 (1.72)	2.00e-06 (2.05)	9.48e-07 (2.00)	6.60e-08 (2.94)	8.50e-09 (2.94)
7	1.26e-05 (1.85)	4.91e-07 (2.03)	2.37e-07 (2.00)	8.44e-09 (2.97)	1.09e-09 (2.97)
8	3.34e-06 (1.92)	1.22e-07 (2.01)	5.92e-08 (2.00)	1.07e-09 (2.98)	1.38e-10 (2.98)
\bar{q}_{num}	1.66	2.19	2.06	2.89	2.80

Table 6.12: A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$
0	7.05e-02	2.51e-02	5.40e-03	9.49e-03	7.49e-04	5.39e-03
1	3.45e-02 (1.03)	7.15e-03 (1.81)	1.08e-03 (2.32)	1.26e-03 (2.92)	1.31e-04 (2.52)	1.08e-03
2	1.54e-02 (1.16)	2.01e-03 (1.83)	2.52e-04 (2.10)	1.84e-04 (2.77)	2.22e-05 (2.56)	2.52e-04
3	6.23e-03 (1.31)	5.59e-04 (1.84)	6.16e-05 (2.03)	2.72e-05 (2.76)	3.48e-06 (2.67)	6.16e-05
4	2.25e-03 (1.47)	1.56e-04 (1.85)	1.53e-05 (2.01)	4.00e-06 (2.76)	5.02e-07 (2.79)	1.53e-05
5	7.29e-04 (1.62)	4.32e-05 (1.85)	3.81e-06 (2.00)	6.25e-07 (2.68)	6.81e-08 (2.88)	3.80e-06
6	2.19e-04 (1.73)	1.20e-05 (1.85)	9.51e-07 (2.00)	1.16e-07 (2.43)	8.90e-09 (2.94)	9.48e-07
7	6.33e-05 (1.79)	3.28e-06 (1.87)	2.38e-07 (2.00)	2.56e-08 (2.18)	1.14e-09 (2.97)	2.37e-07
8	1.79e-05 (1.82)	8.82e-07 (1.89)	5.94e-08 (2.00)	6.16e-09 (2.05)	1.45e-10 (2.97)	5.92e-08
\bar{q}_{num}	1.49	1.85	2.06	2.57	2.79	2.06

Table 6.13: A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

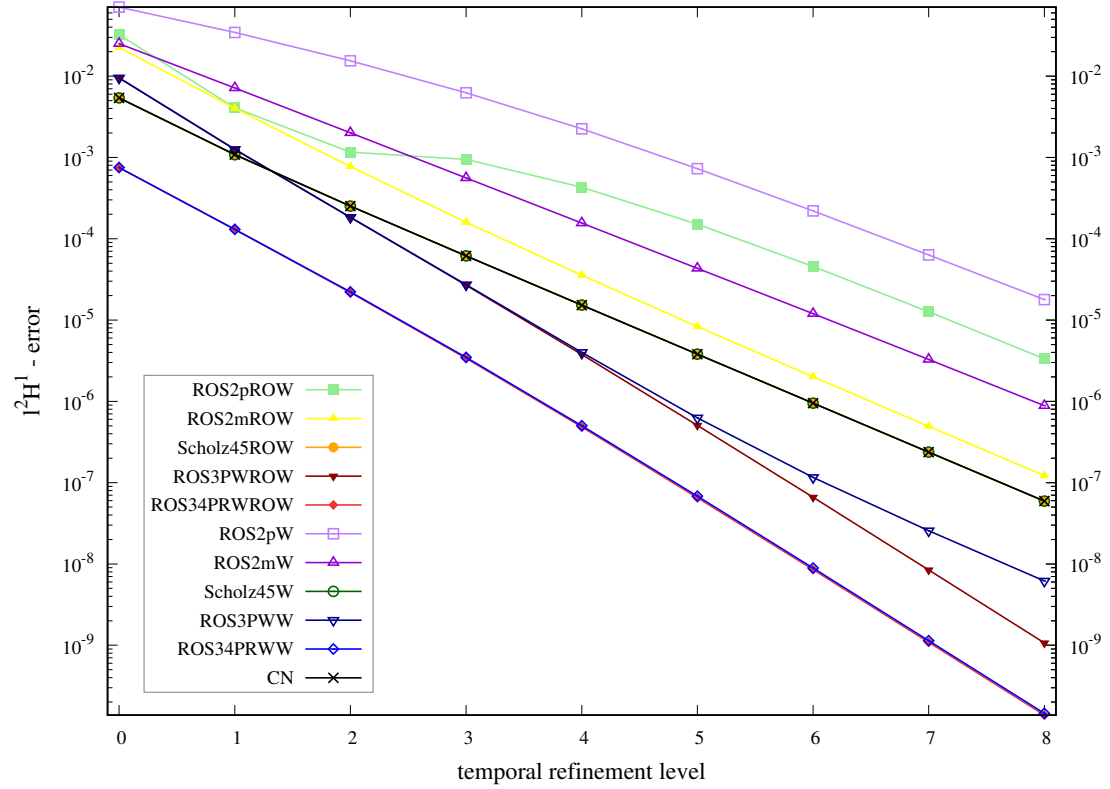


Figure 6.6: A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

Now we slightly change our reaction-diffusion experiments in that we look at a very similar problem with the same exact solution and the same spatial and temporal discretization as before, with the only difference in the problem formulation being that we now look change the diffusion coefficient from $\nu = 1$ to the considerably smaller size of $\nu = 10^{-4}$. Notice, though, that our exact solution is again not affected by that because it is independent of ν . We first look at the resulting $l^\infty L^2$ -errors:

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}
0	8.06e−03	4.46e−03	4.48e−03	1.09e−03	1.93e−03
1	1.74e−03 (2.22)	1.05e−03 (2.08)	1.06e−03 (2.08)	1.38e−04 (2.98)	2.49e−04 (2.96)
2	4.25e−04 (2.03)	2.75e−04 (1.94)	2.77e−04 (1.93)	1.71e−05 (3.02)	3.06e−05 (3.03)
3	1.07e−04 (1.99)	6.88e−05 (2.00)	6.93e−05 (2.00)	2.12e−06 (3.01)	3.79e−06 (3.01)
4	2.69e−05 (1.99)	1.72e−05 (2.00)	1.73e−05 (2.00)	2.64e−07 (3.01)	4.72e−07 (3.01)
5	6.77e−06 (1.99)	4.30e−06 (2.00)	4.33e−06 (2.00)	3.29e−08 (3.00)	5.90e−08 (3.00)
6	1.70e−06 (2.00)	1.08e−06 (2.00)	1.08e−06 (2.00)	4.11e−09 (3.00)	7.36e−09 (3.00)
7	4.25e−07 (1.99)	2.69e−07 (2.00)	2.71e−07 (2.00)	5.13e−10 (3.00)	9.20e−10 (3.00)
8	1.06e−07 (2.00)	6.73e−08 (2.00)	6.76e−08 (2.00)	6.41e−11 (3.00)	1.15e−10 (3.00)
\bar{q}_{num}	2.03	2.00	2.00	3.00	3.00

Table 6.14: A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$
0	4.71e−03	4.70e−03	4.70e−03	1.24e−03	1.98e−03	4.32e−03
1	1.14e−03 (2.04)	1.14e−03 (2.04)	1.14e−03 (2.04)	1.81e−04 (2.77)	2.49e−04 (2.99)	1.07e−03
2	2.95e−04 (1.96)	2.94e−04 (1.96)	2.94e−04 (1.96)	2.89e−05 (2.65)	3.06e−05 (3.02)	2.77e−04
3	7.35e−05 (2.00)	7.33e−05 (2.00)	7.33e−05 (2.00)	6.43e−06 (2.17)	3.80e−06 (3.01)	6.93e−05
4	1.83e−05 (2.00)	1.83e−05 (2.00)	1.83e−05 (2.00)	1.60e−06 (2.01)	4.73e−07 (3.01)	1.73e−05
5	4.59e−06 (2.00)	4.58e−06 (2.00)	4.58e−06 (2.00)	4.00e−07 (2.00)	5.90e−07 (3.00)	4.33e−06
6	1.15e−06 (2.00)	1.14e−06 (2.00)	1.14e−06 (2.00)	1.00e−07 (2.00)	7.37e−09 (3.00)	1.08e−06
7	2.87e−07 (2.00)	2.86e−07 (2.00)	2.86e−07 (2.00)	2.50e−08 (2.00)	9.21e−10 (3.00)	2.71e−07
8	7.16e−08 (2.00)	7.15e−08 (2.00)	7.15e−08 (2.00)	6.25e−09 (2.00)	1.15e−10 (3.00)	6.76e−08
\bar{q}_{num}	2.00	2.00	2.00	2.20	3.00	2.00

Table 6.15: A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

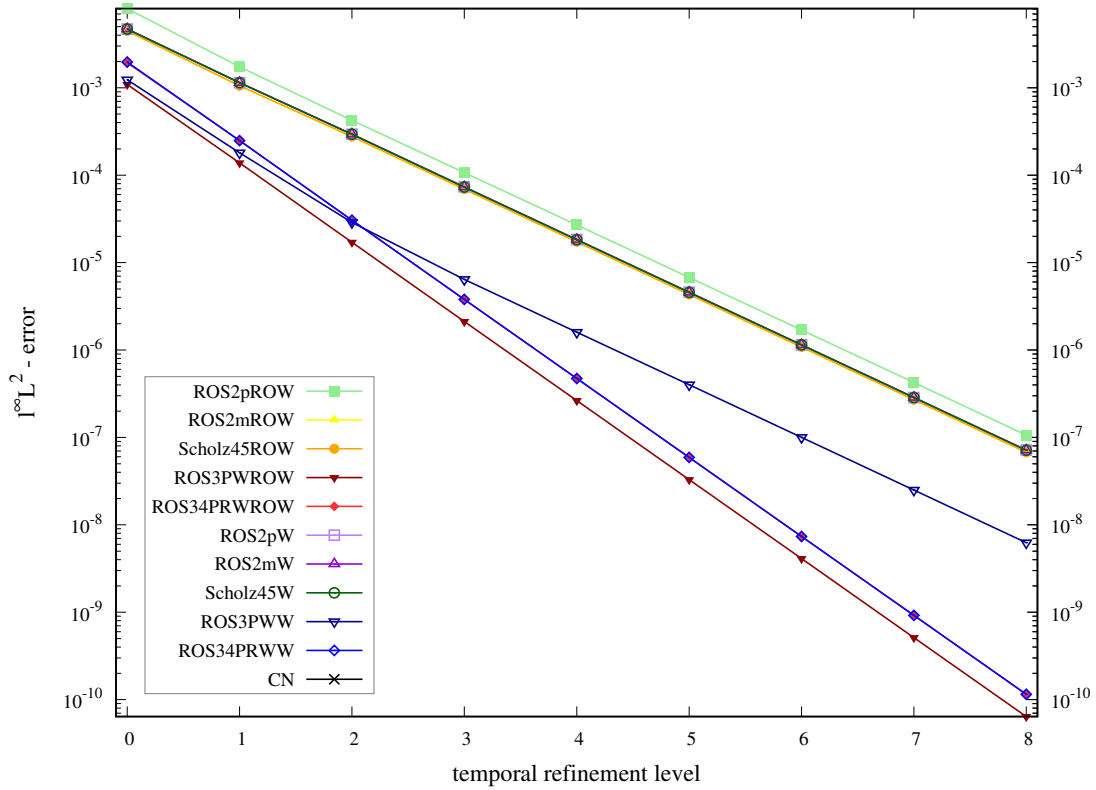


Figure 6.7: A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

In these experiments with diffusion coefficient $\nu = 10^{-4}$ in our reaction-diffusion problem, we observe that both the numerical $l^\infty L^2$ -convergence orders and the numerical $l^2 H^1$ -convergence orders (see below for the $l^2 H^1$ -error data) are for virtually all methods and time step sizes pretty much equal to the corresponding classical convergence orders given in theorem 4.3.1 — with a minor exception being the method ROS3PWW (ROS3PW as W-method with inexact matrix), which has at small time steps a numerical order that is larger than its corresponding classical order 2.

Our interpretation is that decreasing the diffusion coefficient to $\nu = 10^{-4}$ greatly lowers the influence of any spatial derivatives in the equation and thus significantly reduces any temporal stiffness induced by those spatial derivatives. One could say, that our reaction-diffusion equation is almost turned into a non-stiff ODE-like equation, i.e., even simple explicit Runge-Kutta schemes, such as the explicit Euler method, should work as temporal discretization here. This would explain why all our methods behave as in the classical ODE theory and converge with orders that almost exactly match the ones given in theorem 4.3.1.

The crucial difference to the convection-diffusion experiments further above is, that in the reaction-diffusion case here, decreasing the diffusion coefficient to $\nu = 10^{-4}$ lowers the influence of *all* spatial derivatives in the equation, while decreasing the diffusion coefficient to $\nu = 10^{-4}$ in the convection-diffusion equation only influences the diffusive part, leaving the convective part

with its stiffness inducing spatial derivatives untouched.

Below are the $l^2 H^1$ -errors for our reaction-diffusion experiment with $\nu = 10^{-4}$:

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}
0	2.61e-02	1.44e-02	1.45e-02	3.24e-03	5.87e-03
1	5.14e-03 (2.35)	3.48e-03 (2.05)	3.47e-03 (2.06)	3.82e-04 (3.08)	6.87e-04 (3.09)
2	1.22e-03 (2.08)	8.64e-04 (2.01)	8.59e-04 (2.02)	4.67e-05 (3.03)	8.39e-05 (3.03)
3	3.01e-04 (2.02)	2.16e-04 (2.00)	2.14e-04 (2.00)	5.77e-06 (3.02)	1.04e-05 (3.01)
4	7.51e-05 (2.00)	5.39e-05 (2.00)	5.35e-05 (2.00)	7.17e-07 (3.01)	1.29e-06 (3.01)
5	1.88e-05 (2.00)	1.35e-05 (2.00)	1.34e-05 (2.00)	8.94e-08 (3.00)	1.61e-07 (3.00)
6	4.69e-06 (2.00)	3.37e-06 (2.00)	3.34e-06 (2.00)	1.12e-08 (3.00)	2.01e-08 (3.00)
7	1.17e-06 (2.00)	8.41e-07 (2.00)	8.36e-07 (2.00)	1.39e-09 (3.00)	2.51e-09 (3.00)
8	2.94e-07 (2.00)	2.10e-07 (2.00)	2.09e-07 (2.00)	1.74e-10 (3.00)	3.14e-10 (3.00)
\bar{q}_{num}	2.06	2.01	2.01	3.02	3.02

Table 6.16: A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$
0	1.47e-02	1.47e-02	1.47e-02	3.64e-03	5.85e-03	1.41e-02
1	3.60e-03 (2.03)	3.60e-03 (2.03)	3.60e-03 (2.03)	5.87e-04 (2.63)	6.99e-04 (3.07)	3.45e-03
2	8.95e-04 (2.01)	8.95e-04 (2.01)	8.95e-04 (2.01)	1.07e-04 (2.45)	8.56e-05 (3.03)	8.58e-04
3	2.23e-04 (2.00)	2.23e-04 (2.00)	2.23e-04 (2.00)	2.23e-05 (2.27)	1.06e-05 (3.01)	2.14e-04
4	5.58e-05 (2.00)	5.58e-05 (2.00)	5.58e-05 (2.00)	5.05e-06 (2.14)	1.32e-06 (3.01)	5.35e-05
5	1.40e-05 (2.00)	1.40e-05 (2.00)	1.40e-05 (2.00)	1.20e-06 (2.07)	1.64e-07 (3.00)	1.34e-05
6	3.49e-06 (2.00)	3.49e-06 (2.00)	3.49e-06 (2.00)	2.93e-07 (2.04)	2.05e-08 (3.00)	3.34e-06
7	8.72e-07 (1.99)	8.72e-07 (2.00)	8.72e-07 (2.00)	7.22e-08 (2.02)	2.56e-09 (3.00)	8.36e-07
8	2.18e-07 (1.98)	2.18e-07 (2.00)	2.18e-07 (2.00)	1.79e-08 (2.01)	3.20e-10 (3.00)	2.09e-07
\bar{q}_{num}	2.00	2.00	2.00	2.20	3.02	2.01

Table 6.17: A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

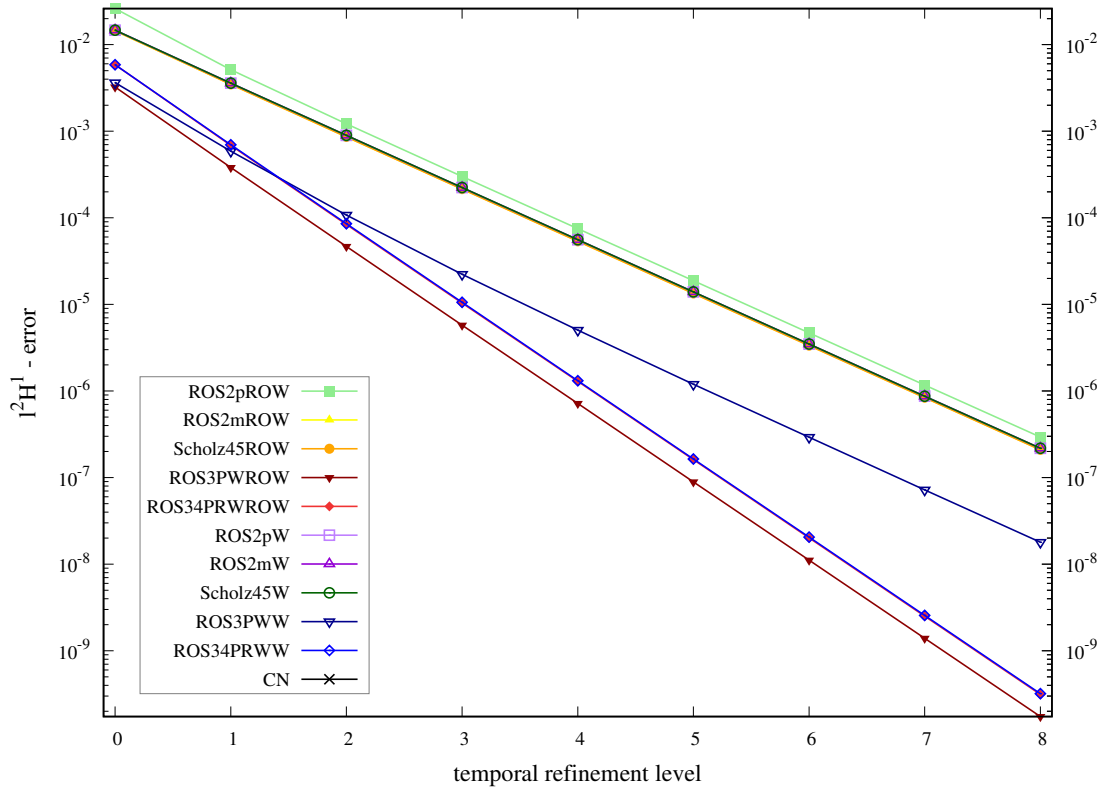


Figure 6.8: A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

6.3 A Navier-Stokes Problem with Known Solution

In this section, we finally study the situation, that the tested equation contains a convective term which is *also* nonlinear — whereas in the previous two sections we examined the difficulties induced by either a linear convective term or a zero-order nonlinear term *separately*.

We begin this section by introducing the problem we look at in detail and then showing how for sufficiently regular data it can be transformed into a problem that matches our theoretical framework.

Throughout this section, let $\Omega := (0, 1)^2$, $\nu \in \mathbb{R}_{>0}$ and as in section 3.4 let

$$H_\sigma(\Omega) := \overline{\{v \in C_0^\infty(\Omega)^2 : \operatorname{div} v = 0\}}^{\|\cdot\|_{L^2}}$$

and equip it with the $L^2(\Omega)^2$ dot product, let

$$V_\sigma(\Omega) := \overline{\{v \in C_0^\infty(\Omega)^2 : \operatorname{div} v = 0\}}^{\|\cdot\|_{H^1}}$$

and equip it with the $H^1(\Omega)^2$ dot product and finally, let

$$Q(\Omega) := \{p \in L^2(\Omega) : (p, 1)_{L^2} = 0\}$$

and equip it with the $L^2(\Omega)^2$ dot product.

We now look at the following problem:

Problem 6.3.1. *Let $v_0 \in H_\sigma(\Omega)$, $g \in C((0, 1]; L^2(\partial\Omega))$ and let $G(t) \in H^{-1}(\Omega)^2$ for all $t \in (0, 1]$. Then we seek a $u = (v, p) \in H^4(0, 1; H^1(\Omega)^2 \times Q(\Omega))$ with*

$$\begin{aligned} \frac{dv}{dt}(t) - \nu \Delta v(t) + (v(t) \cdot \nabla)v(t) + \nabla p(t) &= G(t) && \text{for all } t \in (0, 1], \\ \operatorname{div} v(t) &= 0 && \text{for all } t \in (0, 1], \\ v(t)|_{\partial\Omega} &= g(t) && \text{for all } t \in (0, 1], \\ v(0) &= v_0. \end{aligned}$$

Note that we do not have homogeneous Dirichlet boundary values in this problem. Therefore, it does not directly fit into the theoretical framework from section 3.8. For sufficient regularity of the data, however, one can transform the above problem into the following problem with homogeneous Dirichlet boundary values in the usual way:

Problem 6.3.2. *Let $v_0 \in L^2(\Omega)^2$, $g \in C((0, 1]; L^2(\partial\Omega))$ and let $G(t) \in H^{-1}(\Omega)^2$ for all $t \in (0, 1]$. Now let $v_g \in H^4(0, 1; H^2(\Omega)^2)$ (so $v_g \in C^3([0, 1]; H^2(\Omega)^2)$ by theorem 3.1.2) with $v_g(0)|_{\partial\Omega} = v_0|_{\partial\Omega}$ and $v_g(t)|_{\partial\Omega} = g(t)$ as well as $\operatorname{div} v_g(t) = 0$ for all $t \in (0, 1]$. Then we seek a $u = (\hat{v}, p) \in H^4(0, 1; H_0^1(\Omega)^2 \times Q(\Omega))$ so that*

$$\begin{aligned} \frac{d\hat{v}}{dt}(t) - \nu \Delta \hat{v}(t) + (\hat{v}(t) \cdot \nabla)\hat{v}(t) + (\hat{v}(t) \cdot \nabla)v_g(t) + (v_g(t) \cdot \nabla)\hat{v}(t) + \nabla p(t) \\ = G(t) - \frac{dv_g}{dt}(t) + \nu \Delta v_g(t) - (v_g \cdot \nabla)v_g &&& \text{for all } t \in (0, 1], \\ \operatorname{div} \hat{v}(t) &= 0 && \text{for all } t \in (0, 1], \\ \hat{v}(0) &= v_0 - v_g(0). \end{aligned}$$

We want to mention, that the regularity $v_g \in C^3([0, 1]; H^2(\Omega)^2)$ in this transformed problem makes the operator

$$H^2(\Omega)^2 \cap V_\sigma(\Omega) \ni v \mapsto \nu \mathbb{S}v + P_\sigma[(v_g(t) \cdot \nabla)v + (v \cdot \nabla)v_g(t)] \in H_\sigma(\Omega)$$

for each $t \in [0, 1]$ an Oseen operator as defined in section 3.8.1.

However, to see that problem 6.3.2 fits into the divergence-free framework, which is required for theorems 4.5.1 and 4.5.2, one can first further transform problem 6.3.2 into a problem that does not include a pressure anymore — analogously to proposition 3.8.1 in section 3.8 on the incompressible Navier-Stokes problem — and then define the operator \mathbb{A} as $\mathbb{A} := \nu \mathbb{S}$ and the nonlinearity N as

$$V_\sigma(\Omega) \ni v \mapsto N(v) := -(v \cdot \nabla)v - (v \cdot \nabla)v_g - (v_g \cdot \nabla)v, \cdot)_{L^2} \in [V_\sigma(\Omega)]'$$

The regularity $v_g \in C^3([0, 1]; H^2(\Omega)^2)$ ensures that all requirements of theorems 4.5.1 and 4.5.2 can then be shown with arguments which are very similar to the ones that are used for the original Navier-Stokes problem in subsection 4.5.1.

For implementation purposes, and as we already mentioned in the previous chapter on the finite element method, we are not using the divergence-free framework for our numerical experiments, though.

Furthermore, problem 6.3.1 is usually not solved by first transforming it into problem 6.3.2 and then applying time-stepping methods and spatial discretizations. Rather, the boundary values in problem 6.3.1 are often treated as algebraic conditions that are enforced directly in resulting finite element discretizations. This approach is normally easier to implement and leads to better computational efficiency, thus, we also use it in our experiments here. For details on that way of treating boundary values in Rosenbrock-type methods, we refer to chapter 5 paragraph 1 in [37].

As in the previous experiments, our goal here is to test the performance of Rosenbrock-type methods and study how the error analysis from earlier chapters applies. We do this by choosing an exact solution that fits into the formulation of problem 6.3.1 above and then calculating the right-hand side G and the initial condition v_0 accordingly. In practice, the solution is, of course, not known beforehand. Here, *choosing* the exact solution allows us to measure the errors exactly and it also enables us to pick a solution, which is at each point in time from the finite element space, so that no discretization error in space occurs.

We select the same exact solution that was used in section 6.5 of [50]:

$$(0, 1)^2 \times [0, T] \ni (x, y, t) \mapsto \begin{pmatrix} v_1(x, y, t) \\ v_2(x, y, t) \\ p(x, y, t) \end{pmatrix} := \begin{pmatrix} (y^2 + x) \sin(2\pi t) \\ (x^2 - y) \sin(2\pi t) \\ (x + y - 1)e^{-t} \end{pmatrix} \quad (6.3)$$

We again test all methods from section 4.6 in time-stepping algorithms very similar to the semi-discrete (in time) method 4.5.1. However, problem 6.3.1 does not directly fit into the framework required for method 4.5.1. For example we also have to somehow treat the incompressibility equation in 6.3.1. This means that — other than in the two previous sections with numerical experiments on convection-diffusion and reaction-diffusion problems — merely adjusting method 4.5.1 by having operators and right-hand sides not depend on the semi-discrete solutions from previous time steps and stages, but on the corresponding fully-discrete ones, is not sufficient.

In particular, it is not clear how one needs to handle the pressure and the incompressibility constraint in the time-stepping schemes. Since the full discretization is not the focus of this work however, we do not go into the details there. Roughly speaking, the incompressibility constraint is enforced for each individual stage of the time-stepping scheme, and the pressure is incorporated into the time-stepping scheme alongside the velocity. We essentially use the exact procedure that is demonstrated at the beginning of the paper [67].

In our fully-discrete algorithm, we then obtain, at each stage of our Rosenbrock-type time-stepping schemes, stationary equations similar to 5.12 — with the transport direction \hat{u} in that equation and the right-hand side term f both depending on fully-discrete solutions from previous time steps and stages. That formulation 5.12 covers all our numerical experiments with

Rosenbrock-type methods — in particular it covers all our experiments with different choices for the system matrix in W-methods.

The stationary equations are then discretized via the finite element method that we covered in chapter 5 and — as we already mentioned above — suitable boundary values are enforced directly on that finite element solution.

For comparative purposes we also test the Crank-Nicolson method — using the same ordering of temporal and spatial discretization via Rothe's method as with the Rosenbrock-type methods and also using the same spatial discretization. We solve any nonlinear systems, which arise when using the Crank-Nicolson method here, with Newton's method. For details on discretizing the Navier-Stokes Equations via Rothe's method and the Crank-Nicolson method in that way, see section 3.5.1 in [54].

In the tables and figures below we denote the Crank-Nicolson method in short by CN, the ROS2 method with $\gamma = 1 + \frac{1}{\sqrt{2}}$ by ROS2p and the ROS2 method with $\gamma = 1 - \frac{1}{\sqrt{2}}$ by ROS2m.

We examine all step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$. For the spatial discretization, we cover Ω with a mesh of squares that have sides of length 2^{-l} for some $l \in \{4, 5, 6, 7\}$, and then, based on that mesh, we utilize the LPS-stabilized Q_2/Q_2 -discretization (with suitable stabilization parameters) that we described in section 5.1.2.

For implementation purposes we again keep the spatial discretization the same at each time step — even though changing spatial discretizations between time steps might be desirable for some applications and could also make it easier to test the semi-discrete error estimates from theorems 4.5.1 and 4.5.2.

Now let $(v_{1,m,h}, v_{2,m,h}, p_{m,h})^T$ be the so obtained fully discrete numerical solution — with h being the mesh width and $m \in \{0, 1, \dots, M := \frac{T}{\tau}\}$. Also let $t_m := m\tau$ for all $m \in \{0, 1, \dots, M\}$ as usual. In the following figures and tables of this section, we then refer to the error

$$\max_{0 \leq m \leq M} \left\| \begin{pmatrix} v_{1,m,h} \\ v_{2,m,h} \end{pmatrix} - \begin{pmatrix} v_1(t_m) \\ v_2(t_m) \end{pmatrix} \right\|_{L^2}$$

as the $l^\infty L^2$ -velocity error and to the error

$$\left(\tau \sum_{m=1}^M \left\| \begin{pmatrix} v_{1,m,h} \\ v_{2,m,h} \end{pmatrix} - \begin{pmatrix} v_1(t_m) \\ v_2(t_m) \end{pmatrix} \right\|_{H^1}^2 \right)^{\frac{1}{2}}$$

as the $l^2 H^1$ - velocity error and to the error

$$\left(\tau \sum_{m=1}^M \|p_{m,h} - p(t_m)\|_{L^2}^2 \right)^{\frac{1}{2}}$$

as the $l^2 L^2$ -pressure error.

Even though in section 4.5, no error estimates on the pressure are given as the whole semi-discrete theory is developed in a framework that does not include a pressure, we, of course, look at the pressure error here as well. The particular l^2L^2 -form of the pressure error, as defined above, was chosen to match the usual way of measuring the pressure error — see for example the paper [32], in which the authors also use this particular l^2L^2 -form of the pressure error for their numerical experiments on the instationary incompressible Navier-Stokes equation.

If one wanted to study time-stepping schemes, such as Rosenbrock-type methods, also with the specific intent of reducing the pressure error, one might want to take into account the partial differential *algebraic* nature of problem 6.3.1 — see for example the paper [53] by Rang and Angermann or the papers [49, 50] by Rang.

The respective numerical convergence orders for the above defined errors between successive time step sizes are always denoted by q_{num} , with the average numerical convergence order over all time step sizes, for which feasible solutions were computed, being denoted by \bar{q}_{num} .

For the ROW methods (denoted below with ROW at the end of the name of the method), we have to build at each time step a new stiffness matrix: the Q_2/Q_2 -discretization of the linear terms plus the Q_2/Q_2 -discretization of the Fréchet derivative (at the numerical solution from the last time step) of the convective term.

For the W-methods with inexact stiffness matrices that are tested in most experiments of this section (denoted below with W at the end of the name of the method), we choose to utilize the *same* matrix at each time step: the stiffness matrix, which is used by the ROW methods at the *first* time step. With our specific exact solution and corresponding initial condition, this just means that the approximate stiffness matrix used for these W-methods at each time step is merely the Q_2/Q_2 -discretization of the linear terms.

Towards the end of this section, we will also test W-methods for which we update the stiffness matrix every now and then instead of using the same one at each time step.

In the all experiments of this section, we choose $T = 1$ and we opt for the squares of the spatial mesh to have sides of length 2^{-5} . There are two main sets of experiments, one with kinematic viscosity $\nu = 1$ and one with $\nu = 10^{-4}$. We first look at the results for $\nu = 1$. Here are the $l^\infty L^2$ -velocity errors:

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}
0	5.49e-01	5.48e-01	6.07e-03	9.55e-02	2.57e-03
1	1.42e-01 (1.95)	1.42e-02 (1.95)	3.73e-04 (4.02)	1.08e-02 (3.15)	2.28e-04 (3.49)
2	3.57e-02 (1.99)	3.57e-02 (1.99)	4.61e-05 (3.01)	1.49e-03 (2.85)	1.60e-05 (3.83)
3	8.98e-03 (1.99)	8.97e-03 (1.99)	8.97e-06 (2.36)	1.91e-04 (2.96)	1.16e-06 (3.78)
4	2.25e-03 (2.00)	2.25e-03 (2.00)	2.04e-06 (2.14)	2.40e-05 (2.99)	1.68e-07 (2.79)
5	5.62e-04 (2.00)	5.61e-04 (2.00)	5.08e-07 (2.01)	3.01e-06 (3.00)	2.96e-08 (2.51)
6	1.41e-04 (1.99)	1.40e-04 (2.00)	1.27e-07 (2.00)	3.76e-07 (3.00)	4.72e-09 (2.65)
7	3.55e-05 (1.99)	3.51e-05 (2.00)	3.17e-08 (2.00)	4.69e-08 (3.00)	6.88e-10 (2.78)
8	8.95e-06 (1.99)	8.77e-06 (2.00)	7.93e-09 (2.00)	5.86e-09 (3.00)	9.52e-11 (2.85)
\bar{q}_{num}	1.99	1.99	2.44	2.99	3.09

Table 6.18: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^\infty L^2$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$ q_{num}	$l^\infty L^2$
0	5.49e-01	5.49e-01	1.00e-02	9.56e-02	5.11e-03	8.79e-04
1	1.42e-01 (1.95)	1.42e-01 (1.95)	1.08e-03 (3.21)	1.08e-02 (3.15)	8.91e-04 (2.52)	2.04e-04
2	3.58e-02 (1.99)	3.57e-02 (1.99)	2.63e-04 (2.04)	1.49e-03 (2.85)	1.55e-04 (2.52)	4.57e-05
3	9.00e-03 (1.99)	8.97e-03 (1.99)	6.41e-05 (2.04)	1.91e-04 (2.96)	2.64e-05 (2.55)	9.03e-06
4	2.25e-03 (2.00)	2.25e-03 (2.00)	1.60e-05 (2.00)	2.40e-05 (2.99)	4.12e-06 (2.68)	2.03e-06
5	5.64e-04 (2.00)	5.62e-04 (2.00)	3.99e-06 (2.00)	3.21e-06 (2.90)	6.05e-07 (2.77)	5.07e-07
6	1.41e-04 (2.00)	1.40e-04 (2.00)	9.98e-07 (2.00)	7.45e-07 (2.11)	8.46e-08 (2.84)	1.27e-07
7	3.53e-05 (2.00)	3.51e-05 (2.00)	2.50e-07 (2.00)	1.77e-07 (2.07)	1.14e-08 (2.89)	3.17e-08
8	8.83e-06 (2.00)	8.77e-06 (2.00)	6.24e-08 (2.00)	4.30e-08 (2.04)	1.51e-09 (2.92)	7.93e-09
\bar{q}_{num}	1.99	1.99	2.16	2.64	2.71	2.09

Table 6.19: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^\infty L^2$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

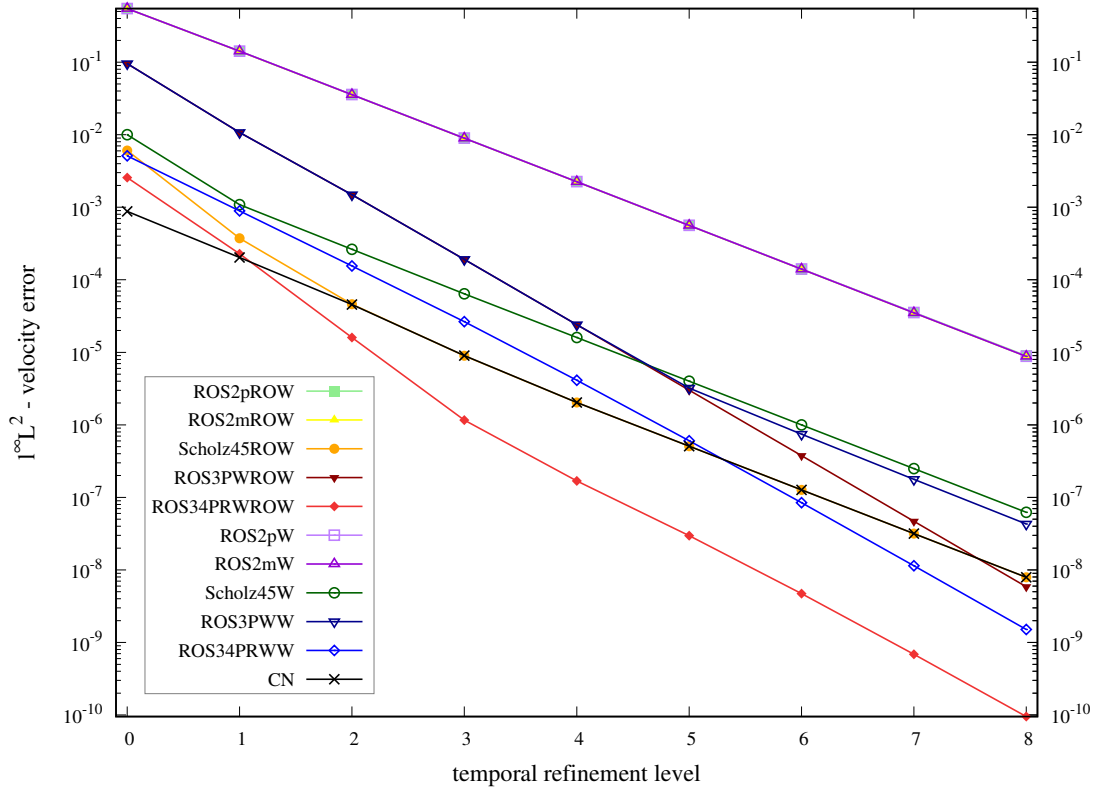


Figure 6.9: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^\infty L^2$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

As we can see, the order 2 methods ROS2p, ROS2m and Scholz45 behave more or less as in the classical ODE theory, i.e., their numerical $l^\infty L^2$ -velocity convergence orders are the same as their corresponding classical ODE orders given in theorem 4.3.1 — a theorem which does not account for effects induced by stiffness, such as order reduction and stability problems. A slight exception can be seen for the method Scholz45 which has at large time steps, and both as ROW method and as W-method with inexact matrix, a numerical order significantly larger than 2.

Also note that all variations of ROS2 (ROS2p and ROS2m as ROW methods and as W-methods with inexact matrices) produce almost the exact same $l^\infty L^2$ -velocity errors for all time step sizes, i.e., using the same stiffness matrix at every step or updating at every step does not seem to make a difference for ROS2p and ROS2m and this specific problem. The method Scholz45ROW, however, does produce significantly smaller $l^\infty L^2$ -velocity errors than the method Scholz45W. Furthermore, at medium to small time steps, Scholz45W produces virtually the same $l^\infty L^2$ -velocity errors as the Crank-Nicolson method.

Now we take a look at the $l^\infty L^2$ -velocity errors produced by our two order 3 methods. The method ROS3PWROW (ROS3PW as ROW method) has for pretty much all time step sizes a numerical $l^\infty L^2$ -velocity convergence order of 3. The method ROS3PWW (our W-method variant of ROS3PW W-method), on the other hand, has for large to medium step sizes a numerical $l^\infty L^2$ -velocity convergence order of 3 and produces almost the same error as the method

ROS3PWROW, but for small step sizes the numerical order drops to 2 and the error becomes a lot worse than for the method ROS3PWROW. These observations about ROS3PW (both as ROW method and as W-method with inexact matrix) again match the statements in theorem 4.3.1, regarding ROS3PW's classical ODE convergence order — as ROS3PW does fulfill all classical ROW method conditions up to order 3 but does **not** fulfill all classical W-method conditions up to order 3.

Looking at the $l^\infty L^2$ -velocity errors for the method ROS34PRWROW (ROS34PRW as ROW method), we see that for large time steps, it has almost order 4, then drops to around order 2.5 for medium sized time steps and finally tends to order 3 for small time steps. ROS34PRW has classical order 3 when used as ROW method for ODEs, and the semi-discrete error estimate 4.5.1 also gives an order 3 error bound. A possible explanation for this behavior of the $l^\infty L^2$ -velocity error produced by ROS34PRWROW in these experiments, is that both the classical ODE error bound from theorem 4.3.1 and the semi-discrete error bound from theorem 4.5.1 might require a fairly small time step size to have a considerable influence on the observed numerical convergence orders.

The method ROS34PRWW (ROS34PRW as W-method with inexact matrix), on the other hand, does not show fast convergence at large time steps (unlike ROS34PRWROW) and produces a significantly larger error than ROS34PRWROW already at those large time steps and also at smaller time steps.

ROS34PRWW starts out with a numerical $l^\infty L^2$ -velocity convergence order of around 2.5 for large time step sizes and then that order steadily tends to 3 for small time steps. ROS34PRW has classical ODE order 3 when used as a W-method with inexact Jacobians, but the semi-discrete W-method error estimate in theorem 4.5.2 (which we already suspect to not be sharp for some methods and problems) only guarantees an order 2 error bound for small enough time steps. It seems very hard to say, which error bounds (classical, semi-discrete in time) have how much of an influence on the $l^\infty L^2$ -velocity error in which regimes of time step sizes here. Maybe the semi-discrete error bounds have more of an influence for large time steps, which would also suggest, though, that the bound from theorem 4.5.2 might again not be sharp in this case — as the observed numerical order is between 2 and 3, i.e., fractional and not 2 as projected in the bound from theorem 4.5.2.

As we already said above in our discussion of the convection-diffusion experiments: fractional temporal orders of convergence for time-stepping schemes applied to parabolic PDEs have been theoretically predicted and numerically observed. For the theoretical analysis of that phenomenon see for example the papers [40, 42, 41] by Lubich and Ostermann. The theorem 4.5.2, which we suspect to not be sharp for ROS34PRWW and the problems we test, is indeed taken from the paper [41]. It is not much of a surprise, though, that theorem 4.5.2 might not be sharp here, as in that theorem and also in the paper [41], the W-methods are only required to have at least classical order 2 — and ROS34PRW has classical order 3 also when used as W-method with inexact Jacobians. Formulating sharp semi-discrete error bounds for some W-methods of higher order could thus be a task for the future.

Overall we see, that for almost all methods and time step sizes, the numerical $l^\infty L^2$ -velocity convergence orders reach at least the semi-discrete order given in theorems 4.5.1 and 4.5.2 — with the exception that we already talked about above being ROS34PRWROW at medium time step sizes.

Judging by the absolute $l^\infty L^2$ -velocity errors in this example, Scholz45 seems to be, for all time step sizes, the best order 2 method — both as ROW method among the order 2 ROW methods and as W-method among the order 2 W-methods. In fact, Scholz45W (our W-method variant of Scholz45) is even considerably better than ROS2pROW (ROS2p as ROW method) and ROS2mROW (ROS2m as ROW method).

Comparing the four stage order 3 method ROS34PRW and the three stage order 3 method ROS3PW, we can conclude that ROS34PRW performs a lot better: ROS34PRWROW and ROS34PRWW both produce significantly smaller $l^\infty L^2$ -velocity errors than ROS3PWROW (which performs a bit better than ROS3PWW) at all time steps.

We also want to mention that ROS3PWROW only reaches $l^\infty L^2$ -velocity errors, which are as low as the ones produced by Scholz45ROW, at the very smallest time steps. ROS3PWW compares even worse with the two stage order 2 method Scholz45, as it only barely manages to outperform Scholz45W and for no time step size it achieves absolute errors as low as the ones produced by Scholz45ROW.

Next, we look at the $l^2 H^1$ -velocity errors. All in all, those errors do not behave much differently than the $l^\infty L^2$ -velocity errors. Hence, we only slightly expand on the above discussion of the $l^\infty L^2$ -velocity errors.

	ROS2pROW		ROS2mROW		Scholz45ROW		ROS3PWROW		ROS34PRWROW	
k	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}
0	8.17e−01		8.22e−01		2.92e−02		1.31e−01		1.61e−02	
1	2.11e−01	(1.95)	2.14e−01	(1.94)	2.31e−03	(3.66)	1.31e−02	(3.32)	1.42e−03	(3.50)
2	5.35e−02	(1.98)	5.46e−02	(1.97)	3.30e−04	(2.80)	1.49e−03	(3.14)	1.15e−04	(3.62)
3	1.36e−02	(1.98)	1.39e−02	(1.97)	6.51e−05	(2.34)	1.79e−04	(3.06)	1.04e−05	(3.47)
4	3.53e−03	(1.94)	3.58e−03	(1.96)	1.51e−05	(2.11)	2.25e−05	(2.99)	1.44e−06	(2.86)
5	9.69e−04	(1.87)	9.28e−04	(1.95)	3.71e−06	(2.03)	2.95e−06	(2.93)	2.49e−07	(2.53)
6	2.84e−04	(1.77)	2.42e−04	(1.94)	9.22e−07	(2.01)	4.01e−07	(2.88)	4.23e−08	(2.55)
7	8.66e−05	(1.71)	6.29e−05	(1.94)	2.30e−07	(2.00)	5.60e−08	(2.84)	6.90e−09	(2.62)
8	2.68e−05	(1.69)	1.63e−05	(1.95)	5.75e−08	(2.00)	7.95e−09	(2.82)	1.09e−09	(2.67)
\bar{q}_{num}	1.86		1.95		2.37		3.00		2.98	

Table 6.20: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 H^1$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW		ROS2mW		Scholz45W		ROS3PWW		ROS34PRWW		CN
k	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$	q_{num}	$l^2 H^1$
0	8.17e-01		8.26e-01		4.84e-02		1.32e-01		2.45e-02		6.90e-03
1	2.11e-01	(1.95)	2.14e-01	(1.95)	5.55e-03	(3.12)	1.41e-02	(3.23)	4.57e-03	(2.42)	1.18e-03
2	5.32e-02	(1.99)	5.43e-02	(1.98)	1.27e-03	(2.13)	1.95e-03	(2.85)	8.47e-04	(2.43)	2.55e-04
3	1.33e-02	(2.00)	1.38e-02	(1.98)	3.11e-04	(2.03)	3.45e-04	(2.50)	1.50e-04	(2.49)	6.05e-05
4	3.34e-03	(2.00)	3.50e-03	(1.97)	7.74e-05	(2.01)	7.16e-05	(2.27)	2.54e-05	(2.57)	1.49e-05
5	8.36e-04	(2.00)	8.96e-04	(1.97)	1.93e-05	(2.00)	1.59e-05	(2.17)	4.10e-06	(2.63)	3.69e-06
6	2.10e-04	(1.99)	2.30e-04	(1.96)	4.83e-06	(2.00)	3.64e-06	(2.12)	6.41e-07	(2.68)	9.21e-07
7	5.29e-05	(1.99)	5.93e-05	(1.96)	1.21e-06	(2.00)	8.59e-07	(2.08)	9.74e-08	(2.72)	2.30e-07
8	1.34e-05	(1.98)	1.52e-05	(1.96)	3.02e-07	(2.00)	2.07e-07	(2.05)	1.44e-08	(2.76)	5.75e-08
\bar{q}_{num}	1.99		1.97		2.16		2.41		2.59		2.11

Table 6.21: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 H^1$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

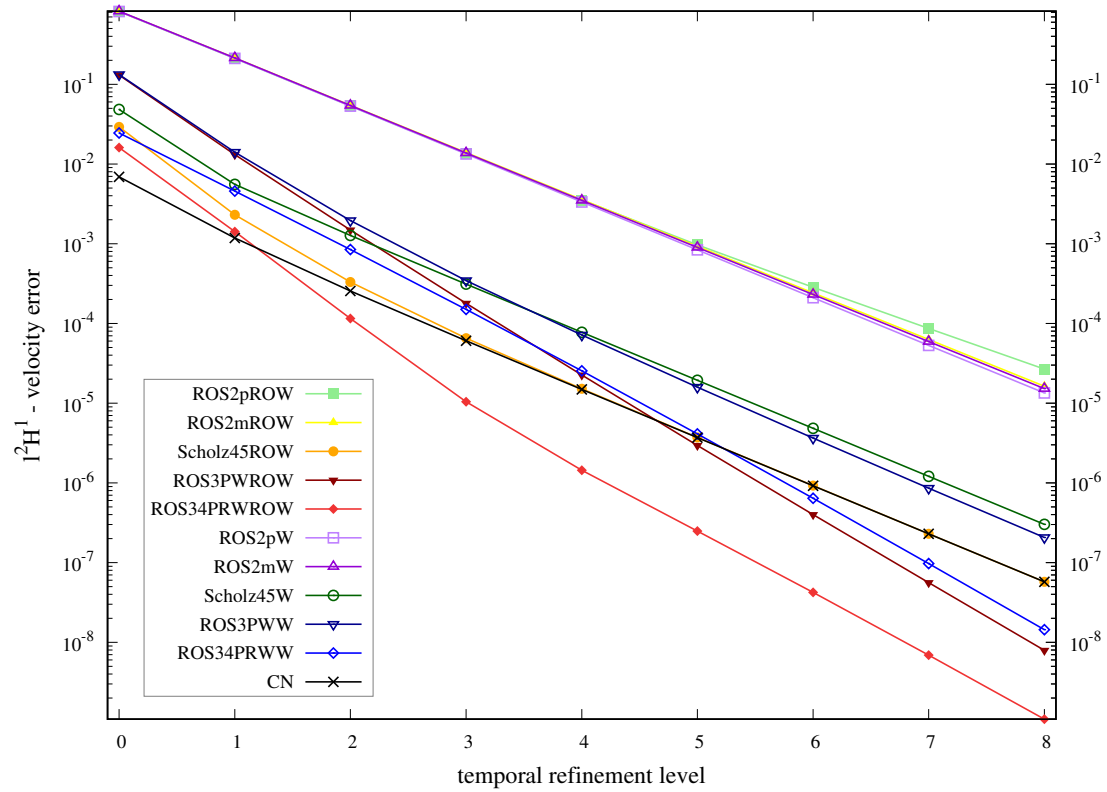


Figure 6.10: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 H^1$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

The $l^2 H^1$ -velocity errors show, as we said above, largely the same behavior as the $l^\infty L^2$ -velocity errors. A notable exception can be seen in the method ROS3PWROW (ROS3PW as ROW method), which in regard to the $l^2 H^1$ -velocity error compares much better with the other methods as opposed to how it compares in regard to the $l^\infty L^2$ -velocity error.

As expected, the absolute value of the $l^2 H^1$ -velocity error is for almost all methods and time step sizes significantly larger than the corresponding $l^\infty L^2$ -velocity error.

Finally, we also look at the corresponding $l^2 L^2$ -pressure errors:

	ROS2pROW		ROS2mROW		Scholz45ROW		ROS3PWROW		ROS34PRWROW	
k	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}
0	1.10e+00		4.63e+00		4.54e-01		4.09e-01		2.11e-01	
1	4.81e-01	(1.19)	2.56e+00	(0.85)	7.08e-02	(2.68)	6.76e-02	(2.60)	2.79e-02	(2.92)
2	2.30e-01	(1.06)	1.31e+00	(0.96)	1.68e-02	(2.07)	1.23e-02	(2.46)	3.53e-03	(2.98)
3	1.14e-01	(1.02)	6.61e-01	(0.99)	4.16e-03	(2.01)	2.49e-03	(2.30)	4.42e-04	(3.00)
4	5.67e-02	(1.00)	3.32e-01	(1.00)	1.04e-03	(2.00)	5.47e-04	(2.19)	5.53e-05	(3.00)
5	2.84e-02	(1.00)	1.66e-01	(1.00)	2.60e-04	(2.00)	1.27e-04	(2.11)	6.92e-06	(3.00)
6	1.42e-02	(1.00)	8.33e-02	(1.00)	6.49e-05	(2.00)	3.05e-05	(2.06)	8.67e-07	(3.00)
7	7.13e-03	(1.00)	4.17e-02	(1.00)	1.62e-05	(2.00)	7.48e-06	(2.03)	1.09e-07	(2.99)
8	3.57e-03	(1.00)	2.08e-02	(1.00)	4.06e-06	(2.00)	1.85e-06	(2.01)	1.37e-08	(2.99)
\bar{q}_{num}	1.03		0.97		2.10		2.22		2.98	

Table 6.22: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 L^2$ -pressure errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW		ROS2mW		Scholz45W		ROS3PWW		ROS34PRWW		CN
k	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}	$l^2 L^2$	q_{num}	$l^2 L^2$
0	5.41e-01		3.98e+00		6.83e-01		5.05e-01		3.37e-01		3.52e-01
1	2.44e-01	(1.15)	2.14e+00	(0.90)	8.51e-02	(3.00)	8.47e-02	(2.58)	8.02e-12	(2.07)	6.88e-02
2	1.22e-01	(1.00)	1.08e+00	(0.99)	2.04e-02	(2.06)	1.80e-02	(2.24)	1.97e-02	(2.03)	1.67e-02
3	6.29e-02	(0.95)	5.38e-01	(1.00)	5.06e-03	(2.01)	4.14e-03	(2.12)	4.88e-03	(2.01)	4.16e-03
4	3.22e-02	(0.96)	2.69e-01	(1.00)	1.26e-03	(2.00)	9.92e-04	(2.06)	1.22e-03	(2.00)	1.04e-03
5	1.64e-02	(0.98)	1.35e-01	(1.00)	3.16e-04	(2.00)	2.42e-04	(2.03)	3.04e-04	(2.00)	2.60e-04
6	8.24e-03	(0.99)	6.74e-02	(1.00)	7.89e-05	(2.00)	5.99e-05	(2.02)	7.58e-05	(2.00)	6.49e-05
7	4.14e-03	(0.99)	3.37e-02	(1.00)	1.97e-05	(2.00)	1.49e-05	(2.01)	1.89e-05	(2.00)	1.62e-05
8	2.07e-03	(1.00)	1.69e-02	(1.00)	4.93e-06	(2.00)	3.71e-06	(2.00)	4.73e-06	(2.00)	4.06e-06
\bar{q}_{num}	1.00		0.99		2.13		2.13		2.01		2.05

Table 6.23: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 L^2$ -pressure errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

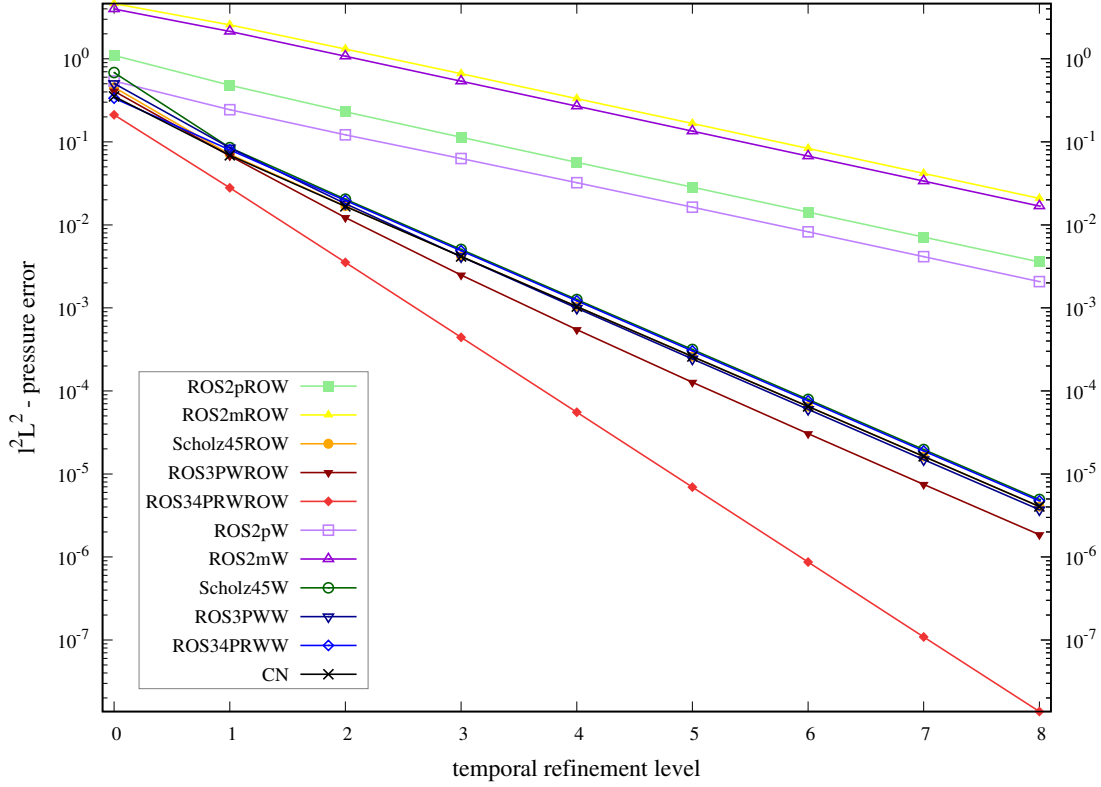


Figure 6.11: A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 L^2$ -pressure error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

As already mentioned above, our theoretical treatment of the temporal discretization of the incompressible Navier-Stokes equations with Rosenbrock-type methods is set in a divergence-free framework without a pressure. Therefore, there are in this work no semi-discrete in time error bounds for the pressure that we could compare our numerical findings to. We again refer to the paper [53] by Rang and Angermann or the papers [49, 50] by Rang for more details on the pressure error produced by ROW methods and W-methods that are applied to the incompressible Navier-Stokes equations.

Looking at the $l^2 L^2$ -pressure errors, we see that any variant of the ROS2 method displays merely a numerical $l^2 L^2$ -pressure convergence order of 1. We found that using ROS2 parameters for the velocity components, but Scholz45 parameters for the pressure components, could raise that convergence order to 2 while maintaining the same velocity errors. We also find it worth mentioning, that in all previous experiments that we examined, ROS2m performed better than ROS2p, but here the $l^2 L^2$ -pressure error is significantly lower for ROS2pROW and ROS2pW than it is for ROS2mROW (which performs marginally better than ROS2mW).

The method Scholz45 (both as ROW method and as W-method with inexact matrix) shows a numerical $l^2 L^2$ -pressure convergence order of 2 — as does the method ROS3PW (again both as ROW method and as W-method). The ROS34PRW is notable in that its W-method variant also has a numerical $l^2 L^2$ -pressure convergence order of merely 2, while its ROW method variant

is in fact the only method we tested that maintains order 3 convergence for the pressure.

Comparing for each method the ROW method variant with the W-method variant, we see that for almost all methods, the W-method variant performs only slightly worse than the ROW method variant. The one exception is the method ROS34PRW, which we just discussed — with ROS34PRWROW having a higher numerical l^2L^2 -pressure convergence order than ROS34PRWW and thus for most time step sizes also a significantly lower absolute l^2L^2 -pressure error.

Now we slightly change our Navier-Stokes experiments in that we look at a very similar problem with the same exact solution and the same spatial and temporal discretization as before, with the only difference in the problem formulation being that we now change the kinematic viscosity from $\nu = 1$ to the considerably smaller size of $\nu = 10^{-4}$. Once again, we want to mention, though, that our exact solution is not affected by that since it does not depend on ν . We first look at the resulting $l^\infty L^2$ -velocity errors:

	ROS2pROW		ROS2mROW		Scholz45ROW		ROS3PWROW		ROS34PRWROW	
k	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}
0	5.49e−01		8.05e+04		4.67e−02		9.75e−02		1.17e+11	
1	1.42e−01	(1.95)	1.48e−01		5.30e−03	(3.14)	1.08e−02	(3.18)	5.56e−03	
2	3.66e−02	(1.96)	3.57e−02	(2.05)	1.24e−03	(2.10)	1.49e−03	(2.85)	2.52e−04	(4.46)
3	9.49e−03	(1.95)	8.97e−03	(1.99)	3.13e−04	(1.98)	1.90e−04	(2.96)	3.19e−05	(2.98)
4	2.45e−03	(1.95)	2.24e−03	(2.00)	7.93e−05	(1.98)	2.39e−05	(2.99)	4.15e−06	(2.94)
5	6.28e−04	(1.97)	5.61e−04	(2.00)	1.99e−05	(1.99)	3.00e−06	(3.00)	5.29e−07	(2.97)
6	1.59e−04	(1.98)	1.40e−04	(2.00)	4.99e−06	(2.00)	3.75e−07	(3.00)	6.67e−08	(2.99)
7	3.98e−05	(1.99)	3.51e−05	(2.00)	1.25e−06	(2.00)	4.69e−08	(3.00)	8.38e−09	(2.99)
8	9.98e−06	(2.00)	8.77e−06	(2.00)	3.12e−07	(2.00)	5.86e−09	(3.00)	1.05e−09	(3.00)
\bar{q}_{num}	1.97		2.01		2.15		3.00		3.19	

Table 6.24: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW		ROS2mW		Scholz45W		ROS3PWW		ROS34PRWW		CN
k	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$	q_{num}	$l^\infty L^2$
0	1.57e+08		4.57e+09		2.24e+15		9.88e+08		1.04e+08		5.47e+06
1	7.25e+09		2.06e+10		3.06e+09		1.30e+11		1.34e+13		5.30e-03
2	5.40e+09		9.03e+07		1.89e+11		4.55e+13		2.55e+08		1.29e-03
3	1.90e+09		9.78e+11		6.47e+08		3.34e+08		3.13e-05		3.21e-04
4	2.75e-03		2.36e-03		8.82e-04		7.07e-04		3.34e-06 (3.23)		8.00e-05
5	5.89e-04 (2.22)		5.89e-04 (2.00)		1.12e-04 (2.98)		7.84e-05 (3.17)		4.04e-07 (3.05)		2.00e-05
6	1.47e-04 (2.00)		1.47e-04 (2.00)		2.79e-05 (2.00)		1.96e-05 (2.00)		4.97e-08 (3.02)		5.00e-06
7	3.68e-05 (2.00)		3.68e-05 (2.00)		6.98e-06 (2.00)		4.90e-06 (2.00)		6.17e-09 (3.01)		1.25e-06
8	9.20e-06 (2.00)		9.20e-06 (2.00)		1.74e-06 (2.00)		1.22e-06 (2.00)		7.68e-10 (3.01)		3.12e-07
\bar{q}_{num}	2.06		2.00		2.25		2.29		3.06		2.01

Table 6.25: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

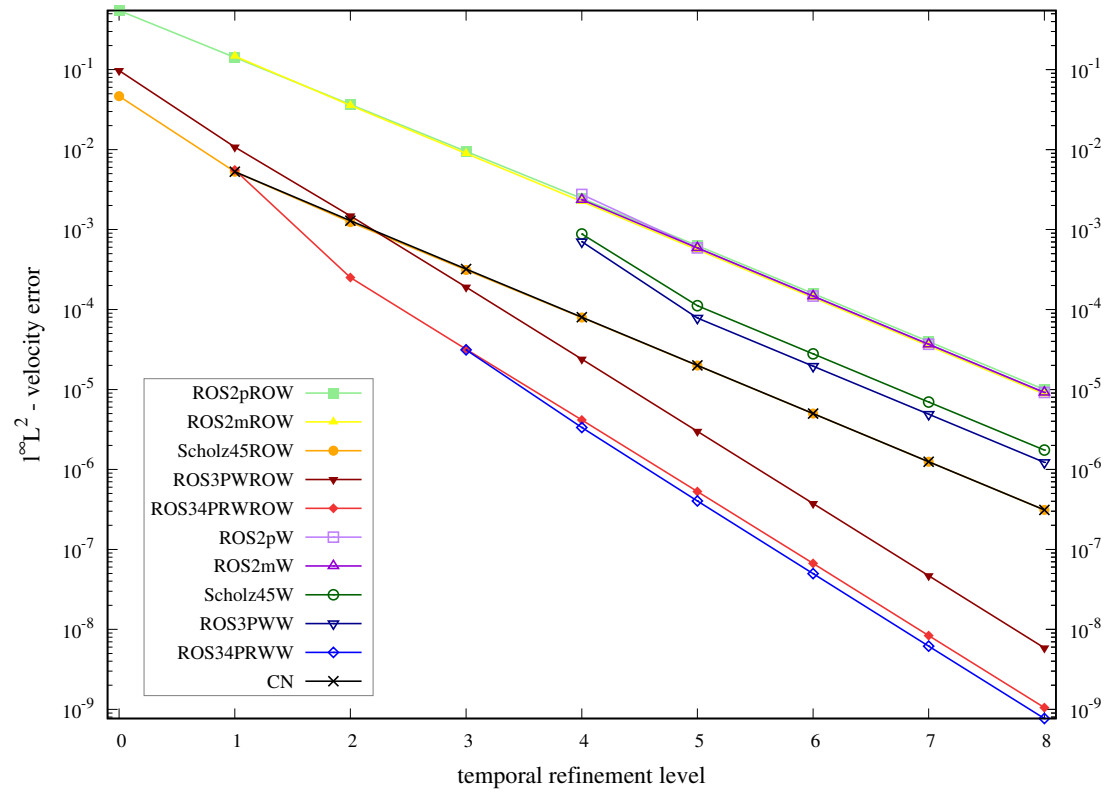


Figure 6.12: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

In these experiments with small kinematic viscosity we notice a similar situation as we saw in section 6.1 in our experiments on a convection-diffusion problem with small diffusion coefficient: If the time step is too large, then our W-methods with inexact stiffness matrices encounter stability issues and do not produce feasible solutions. For the time step sizes that do yield viable solutions, the numerical $l^\infty L^2$ -velocity convergence orders are for all of our W-methods nearly the same as the corresponding classical ODE orders given in theorem 4.3.1.

The ROW methods, on the other hand, work for almost all time step sizes (with the exception being the very largest time step, for which only ROS2mROW and ROS34PRWROW fail) and their numerical $l^\infty L^2$ -velocity convergence order is in all time step size regimes very close to their corresponding classical ODE order given in theorem 4.3.1.

Looking at the absolute $l^\infty L^2$ -velocity errors in the convection-dominated case ($\nu = 10^{-4}$), we see that when they produce viable solutions, all ROS2 variants display nearly the same errors. Remarkably, those errors are also almost exactly the same as in the standard case ($\nu = 1$).

For the methods Scholz45 and ROS3PW, the $l^\infty L^2$ -velocity errors in the convection-dominated case are significantly smaller for their ROW method variants than for their W-method variants. Furthermore, the methods Scholz45ROW (which in the convection-dominated case shows essentially the same $l^\infty L^2$ -velocity errors as the Crank-Nicolson method), Scholz45W and ROS3PW all generate considerably lower $l^\infty L^2$ -velocity errors in the standard case than in the convection-dominated case. For the method ROS3PWROW those errors are more or less equal.

The $l^\infty L^2$ -velocity errors for the viable solutions produced by ROS34PRWW in the convection-dominated case are remarkable here in two ways: Firstly, those errors are actually a bit *smaller* than the corresponding errors displayed by ROS34PRWROW in the convection-dominated case. Secondly, those errors for the viable solutions produced by ROS34PRWW in the convection-dominated case are also smaller than the corresponding errors generated by ROS34PRWW in the standard case.

Judging by the $l^\infty L^2$ -velocity errors in the convection-dominated case, the method ROS34PRWROW seems to perform best, as it is both stable and fast. If small time steps may be used, the method ROS34PRWW is a very good option as well for this example problem. The best performing order 2 method is again Scholz45 — both as ROW method and as W-method with inexact matrix.

We now present the $l^2 H^1$ -velocity error results for our convection-dominated Navier-Stokes problem.

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}
0	4.05e+00	4.09e+06	2.13e+00	1.43e+00	5.36e+12
1	1.27e+00 (1.67)	2.78e+00	3.37e-01 (2.66)	2.17e-01 (2.71)	2.57e-01
2	4.54e-01 (1.48)	1.47e-01 (4.24)	7.24e-02 (2.22)	2.62e-02 (3.05)	1.62e-02 (3.98)
3	1.66e-01 (1.45)	2.37e-02 (2.64)	1.67e-02 (2.12)	2.91e-03 (3.17)	1.80e-03 (3.17)
4	5.28e-02 (1.66)	5.01e-03 (2.24)	4.01e-03 (2.06)	3.28e-04 (3.15)	2.12e-04 (3.09)
5	1.48e-02 (1.84)	1.19e-03 (2.07)	9.85e-04 (2.02)	3.93e-05 (3.06)	2.58e-05 (3.04)
6	3.85e-03 (1.94)	2.93e-04 (2.02)	2.44e-04 (2.01)	4.85e-06 (3.02)	3.18e-06 (3.02)
7	9.76e-04 (1.98)	7.30e-05 (2.01)	6.08e-05 (2.01)	6.05e-07 (3.00)	3.96e-07 (3.01)
8	2.45e-04 (1.99)	1.82e-05 (2.00)	1.52e-05 (2.00)	7.57e-08 (3.00)	4.94e-08 (3.00)
\bar{q}_{num}	1.75	2.46	2.14	3.02	3.19

Table 6.26: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$
0	9.53e+09	2.13e+11	1.04e+17	4.20e+10	6.99e+09	2.80e+08
1	4.39e+11	1.42e+12	2.24e+11	7.87e+12	1.21e+15	2.51e-01
2	3.87e+11	7.33e+09	1.52e+13	3.40e+15	2.04e+10	6.29e-02
3	1.10e+11	6.05e+13	3.81e+10	2.02e+10	9.85e-04	1.57e-02
4	1.07e-01	6.76e-02	3.58e-02	2.78e-02	1.07e-04 (3.20)	3.90e-03
5	9.92e-03 (3.42)	9.74e-03 (2.80)	4.39e-03 (3.03)	2.68e-03 (3.37)	1.26e-05 (3.08)	9.73e-04
6	2.48e-03 (2.00)	2.43e-03 (2.00)	1.10e-03 (2.00)	6.70e-04 (2.00)	1.54e-06 (3.03)	2.43e-04
7	6.19e-04 (2.00)	6.07e-04 (2.00)	2.74e-04 (2.00)	1.67e-04 (2.00)	1.91e-07 (3.01)	6.07e-05
8	1.55e-04 (2.00)	1.52e-04 (2.00)	6.84e-05 (2.00)	4.18e-05 (2.00)	2.37e-08 (3.01)	1.52e-05
\bar{q}_{num}	2.36	2.20	2.26	2.34	3.07	2.00

Table 6.27: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

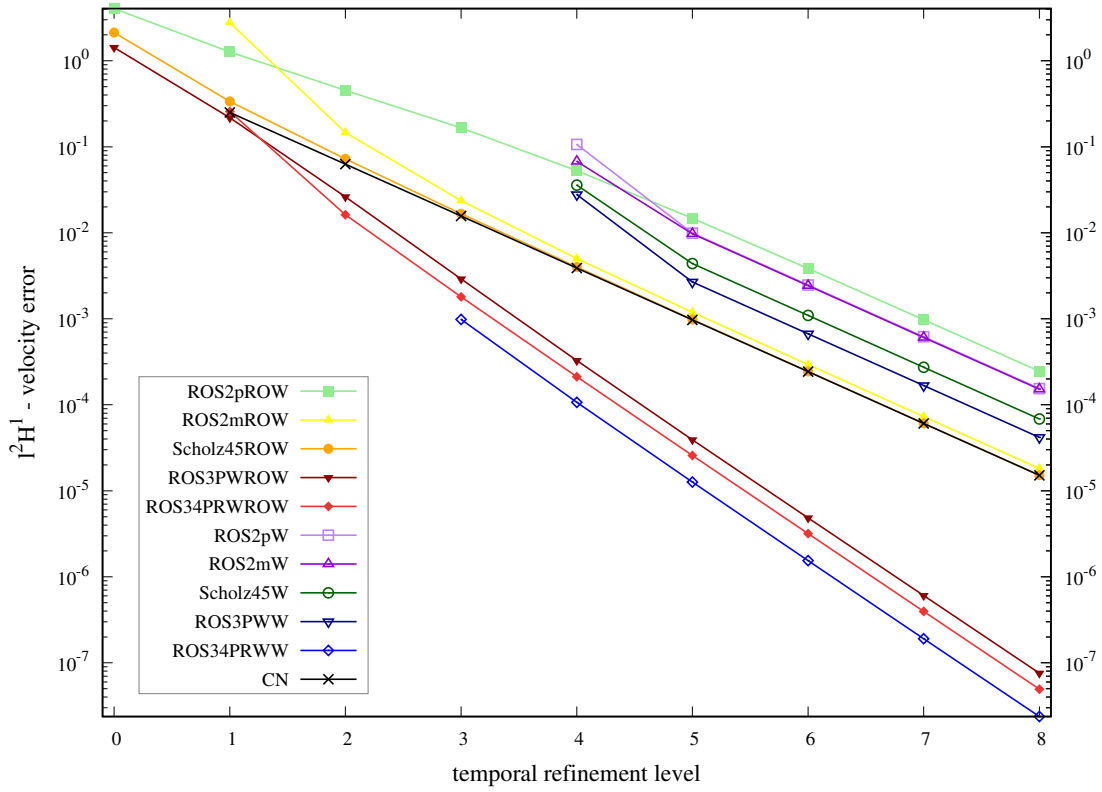


Figure 6.13: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

In the convection-dominated case, these $l^2 H^1$ -velocity errors again do not add a lot more insight to what we gathered above from the $l^\infty L^2$ -errors. There are few things we want to mention, though.

Namely, in regard to the absolute $l^2 H^1$ -velocity error, the methods ROS2mROW and ROS3PWROW compare much better with the other methods as opposed to how they compare in regard to the $l^\infty L^2$ -velocity error.

Moreover, for the method ROS34PRWW, which in the convection-dominated case actually generated a smaller $l^\infty L^2$ -velocity error than the method ROS34PRWROW, that gap now even widens a bit when looking at the $l^2 H^1$ -velocity error. However, in contrast to the $l^\infty L^2$ -velocity error, the $l^2 H^1$ -velocity error is not smaller for ROS34PRWW in the convection-dominated case than in the standard case.

Here are the $l^2 L^2$ -pressure errors for our experiments with the convection-dominated problem:

	ROS2pROW	ROS2mROW	Scholz45ROW	ROS3PWROW	ROS34PRWROW
k	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}
0	1.11e+00	9.62e+04	4.51e-01	4.22e-01	1.22e+11
1	4.85e-01 (1.19)	2.59e+00	7.12e-02 (2.66)	6.79e-02 (2.64)	2.82e-02
2	2.33e-01 (1.06)	1.33e+00 (0.97)	1.69e-02 (2.07)	1.24e-02 (2.45)	3.56e-03 (2.99)
3	1.15e-01 (1.02)	6.66e-01 (0.99)	4.19e-03 (2.01)	2.52e-03 (2.30)	4.47e-04 (2.99)
4	5.73e-02 (1.00)	3.34e-01 (1.00)	1.04e-03 (2.00)	5.53e-04 (2.19)	5.59e-05 (3.00)
5	2.86e-02 (1.00)	1.67e-01 (1.00)	2.61e-04 (2.00)	1.28e-04 (2.11)	6.99e-06 (3.00)
6	1.43e-02 (1.00)	8.34e-02 (1.00)	6.52e-05 (2.00)	3.07e-05 (2.06)	8.74e-07 (3.00)
7	7.16e-03 (1.00)	4.17e-02 (1.00)	1.63e-05 (2.00)	7.52e-06 (2.03)	1.09e-07 (3.00)
8	3.58e-03 (1.00)	2.09e-02 (1.00)	4.07e-06 (2.00)	1.86e-06 (2.02)	1.37e-08 (3.00)
\bar{q}_{num}	1.03	0.99	2.09	2.22	3.00

Table 6.28: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 L^2$ -pressure errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

	ROS2pW	ROS2mW	Scholz45W	ROS3PWW	ROS34PRWW	CN
k	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$
0	1.90e+07	7.78e+08	4.56e+14	1.07e+08	2.38e+07	4.08e+06
1	1.25e+09	1.11e+10	1.21e+09	2.74e+10	5.96e+12	6.95e-02
2	1.52e+09	1.48e+08	7.27e+10	2.65e+13	2.95e+08	1.69e-02
3	8.06e+08	8.36e+11	4.80e+08	5.43e+08	4.85e-03	4.18e-03
4	8.69e-03	2.68e-01	1.28e-03	1.00e-03	1.21e-03 (2.00)	1.04e-03
5	4.25e-03 (1.03)	1.34e-01 (1.00)	3.21e-04 (2.00)	2.45e-04 (2.03)	3.02e-04 (2.00)	2.61e-04
6	2.11e-03 (1.00)	6.71e-02 (1.00)	8.02e-05 (2.00)	6.05e-05 (2.02)	7.56e-05 (2.00)	6.52e-05
7	1.06e-03 (1.00)	3.35e-02 (1.00)	2.00e-05 (2.00)	1.50e-05 (2.01)	1.89e-05 (2.00)	1.63e-05
8	5.30e-04 (1.00)	1.68e-02 (1.00)	5.01e-06 (2.00)	3.75e-06 (2.00)	4.72e-06 (2.00)	4.07e-06
\bar{q}_{num}	1.01	1.00	2.00	2.02	2.00	2.01

Table 6.29: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 L^2$ -pressure errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$

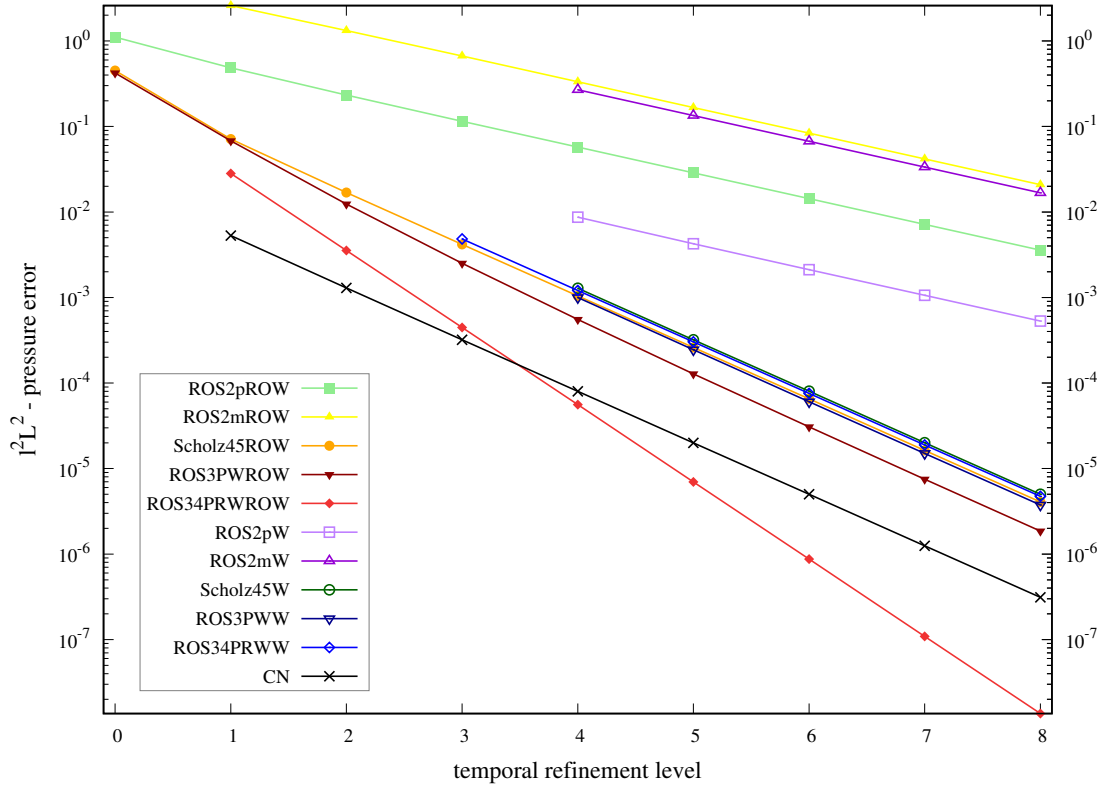


Figure 6.14: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 L^2$ -pressure error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

Since in the convection-dominated case ($\nu = 10^{-4}$) the W-methods with inexact stiffness matrices do not produce feasible solutions for large time step sizes, we obviously do not obtain a pressure for those time step sizes either.

In comparison with the standard case ($\nu = 1$), we can see that almost all methods generate more or less the same absolute $l^2 L^2$ -pressure errors in the convection-dominated case (when viable solutions are produced) — with the exception being the method ROS2pW, which at small time steps actually generates significantly *lower* $l^2 L^2$ -pressure errors in the convection-dominated case than in the standard case.

In all previous experiments, we tested W-methods that used one and the same stiffness matrix for *every* time step. Now we want to employ Scholz45 (so far the best performing order 2 method) and ROS34PRW (so far the best performing order 3 method) as W-methods in a different way: Instead of building a stiffness matrix only once, at the first time step and then using that same matrix at each following time step, we now build a matrix at the first step and then update it every 5th/20th/80th step. In the tables and figures below, these W-method variants will be denoted with W5/W20/W80 at the end of the name of the methods — with the standard ROW methods still being denoted with ROW at the end of their name and the W-methods that use only one and the same matrix each time step still being denoted with W at the end of their name.

In the new W-methods with the occasional matrix update, the computing cost of the algorithm is certainly increased compared to the W-methods that only require one stiffness matrix to be build. Our hope is, of course, that they are also more stable and/or accurate.

In the following experiments, we look at the same convection-dominated Navier-Stokes problem with the same exact solution and the same spatial and temporal discretization techniques as before, with the only difference in the temporal discretization being that we use the aforementioned W-methods with the occasional matrix updates.

In the graphs below, we compare our findings with the previously presented results for Scholz45ROW/W and ROS34PRWROW/W. To keep the presentation a bit shorter and because no substantial insight is lost that way, we present for the velocity only the $l^2 H^1$ -errors, i.e., we omit the $l^\infty L^2$ -velocity errors.

	Scholz45W80	Scholz45W20	Scholz45W5	ROS34PRWW80	ROS34PRWW20	ROS34PRWW5
k	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}	$l^2 H^1$ q_{num}
0	1.04e+17	1.04e+17	1.04e+17	6.99e+09	6.99e+09	6.99e+09
1	2.24e+11	2.24e+11	6.19e+16	1.21e+15	1.21e+15	1.21e+15
2	1.52e+13	1.52e+13	4.94e+15	2.04e+10	2.04e+10	1.12e+14
3	3.81e+10	1.67e+16	2.47e-02	9.85e-04	9.85e-04	2.59e-03
4	3.58e-02	1.33e-02	4.40e-03 (2.49)	1.07e-04 (3.20)	1.07e-03 (-0.12)	2.14e-04 (3.60)
5	4.39e-03 (3.03)	1.51e-03 (3.15)	1.01e-03 (2.12)	1.26e-05 (3.08)	4.98e-05 (4.42)	2.53e-05 (3.08)
6	8.07e-04 (2.44)	2.71e-04 (2.47)	2.46e-04 (2.04)	1.82e-05 (-0.53)	3.33e-06 (3.90)	3.12e-06 (3.02)
7	9.32e-05 (3.11)	6.27e-05 (2.11)	6.11e-05 (2.01)	7.95e-07 (4.52)	3.88e-07 (3.10)	3.88e-07 (3.01)
8	1.69e-05 (2.47)	1.54e-05 (2.03)	1.52e-05 (2.01)	5.21e-08 (3.93)	4.82e-08 (3.10)	4.84e-08 (3.00)
\bar{q}_{num}	2.76	2.44	2.13	2.84	2.86	3.14

Table 6.30: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity errors for Scholz45 and ROS34PRW (system matrix update only every 80th/20th/5th time step), with temporal refinement level $k \in \{0, \dots, 8\}$, $5 \cdot 2^k$ time steps and time step size $0.2 \cdot 2^{-k}$

Regarding the stability of the methods, we hardly see a change here. Only the method Scholz45W5 (matrix update every 5th step) slightly increases the maximally admissible time step size compared to the method Scholz45W (same matrix every step): among the tested time step sizes, the largest one which still leads to a feasible solution is $\tau = 0.2 \cdot 2^{-3}$ for Scholz45W5 and $\tau = 0.2 \cdot 2^{-4}$ for Scholz45W20, Scholz45W80 and also Scholz45W.

Regarding ROS34PRW, we see that all its tested W-method variants with inexact matrix have the same maximally admissible time step size among those step sizes we used. Apparently, updating the stiffness matrix every now and then, as opposed to only using the same one at every step, does not substantially raise the stability of ROS34PRW.

Below are the same findings in the form of a graph:

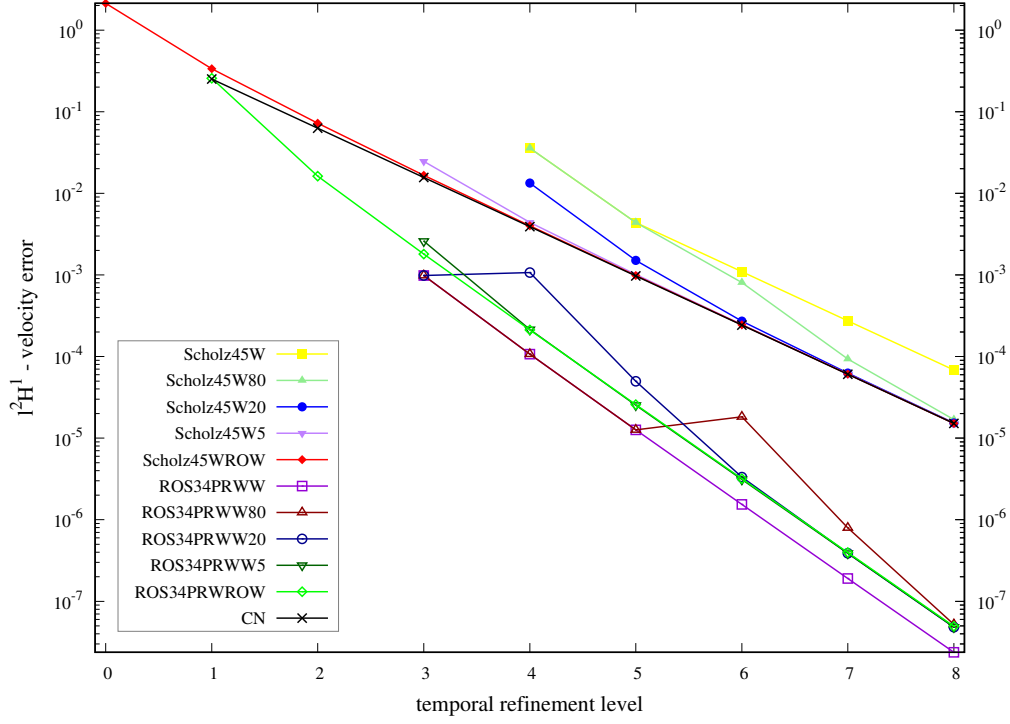


Figure 6.15: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution — testing occasional updating of the stiffness matrix in Scholz45 and ROS34PRW: $l^2 H^1$ -velocity error versus temporal refinement level k (with $5 \cdot 2^k$ time steps and the time step size $0.2 \cdot 2^{-k}$ in the experiments)

In contrast to the stability of the W-methods we tested, the absolute $l^2 H^1$ -velocity errors generated by those methods are significantly impacted by the varying update-frequency of the stiffness matrix.

In our experiments with the Scholz45 variants, we make the following, more or less expected, observations: The more frequent the matrix update is, the larger are the time step sizes for which we see a substantial decrease in the $l^2 H^1$ -velocity errors compared to those generated by Scholz45W (the W-method variant with one and the same stiffness matrix at every time step). For the smallest time steps, all Scholz45 W-method variants with occasional matrix update produce essentially the same $l^2 H^1$ -velocity error as the method Scholz45ROW. Only the method Scholz45W generates a significantly larger error than Scholz45ROW at the smallest time step.

The $l^2 H^1$ -velocity errors generated by the ROS34PRW W-method variants seem a bit peculiar — as is displayed in the above graph and table. For some step size regimes, not updating the matrix at all generates smaller errors than updating the matrix *very often*, which in turn generates smaller errors than updating the matrix *rarely*. Generally, among the ROS34PRW W-methods we tested, the one which uses the same stiffness matrix at every step still generates the smallest $l^2 H^1$ -velocity errors — it even outperforms ROS34PRWROW for medium and small time steps.

We now take a look at the corresponding $l^2 L^2$ -pressure errors:

	Scholz45W80	Scholz45W20	Scholz45W5	ROS34PRWW80	ROS34PRWW20	ROS34PRWW5
k	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}	$l^2 L^2$ q_{num}
0	4.56e+14	4.56e+14	4.56e+14	2.38e+07	2.38e+07	2.38e+07
1	1.21e+09	1.21e+09	4.38e+15	5.96e+12	5.96e+12	5.96e+12
2	7.27e+10	7.27e+10	1.10e+15	2.95e+08	2.95e+08	7.24e+11
3	4.80e+08	3.44e+14	4.59e-03	4.85e-03	4.85e-03	1.52e-03
4	1.28e-03	1.68e-03	1.06e-03 (2.11)	1.21e-03 (2.00)	9.04e-04 (2.42)	1.85e-04 (3.04)
5	3.21e-04 (2.00)	3.31e-04 (2.35)	2.62e-04 (2.02)	3.02e-04 (2.00)	8.72e-05 (3.37)	2.29e-05 (3.01)
6	1.05e-04 (1.61)	7.17e-05 (2.21)	6.53e-05 (2.01)	5.56e-05 (2.44)	1.04e-05 (3.06)	2.85e-06 (3.00)
7	2.07e-05 (2.35)	1.69e-05 (2.08)	1.63e-05 (2.00)	5.33e-06 (3.38)	1.29e-06 (3.02)	3.56e-07 (3.00)
8	4.49e-06 (2.21)	4.15e-06 (2.03)	4.08e-06 (2.00)	6.36e-07 (3.07)	1.61e-07 (3.00)	4.45e-08 (3.00)
\bar{q}_{num}	2.04	2.17	2.03	2.58	2.98	3.01

Table 6.31: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 L^2$ -pressure errors for Scholz45 and ROS34PRW (system matrix update only every 80th/20th/5th time step), with temporal refinement level $k \in \{0, \dots, 8\}$, $5 \cdot 2^k$ time steps and time step size $0.2 \cdot 2^{-k}$

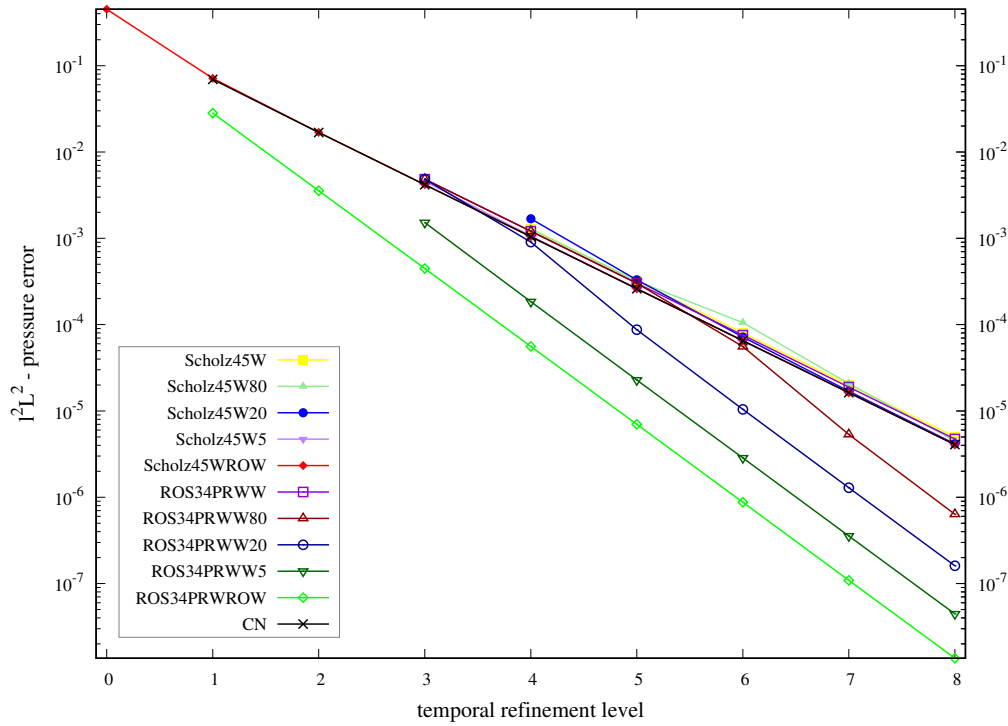


Figure 6.16: A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution — testing occasional updating of the system matrix in Scholz45 and ROS34PRW: $l^2 L^2$ -pressure error versus temporal refinement level k (with $5 \cdot 2^k$ time steps and the time step size $0.2 \cdot 2^{-k}$ in the experiments)

Here we observe a very straightforward result. For the W-method variants of Scholz45, the frequency with which the stiffness matrix is updated has almost no influence on the l^2L^2 -pressure errors. The W-method variants of ROS34PRW, on the other hand, produce progressively smaller pressure errors with an increasing update frequency of the stiffness matrix. All our ROS34PRW W-method variants with occasional matrix update eventually reach a numerical l^2L^2 -pressure convergence order of 3 for small time steps — with ROS34PRWW (same matrix at every time step) being the only ROS34PRW W-method in our experiments that remains at a numerical l^2L^2 -pressure convergence order of merely 2, even for small time steps.

This certainly creates a dilemma in regard to choosing the best ROS34PRW variant for the convection-dominated Navier-Stokes problem we are currently looking at: If one wanted the smallest pressure error, one had to choose ROS34PRWW5 or even ROS34PRWWO. If, however, one wanted to minimize the velocity error at medium and small time steps, one had to choose ROS34PRWW (same matrix at every time step) as shown above in the table and graph on the l^2H^1 -velocity errors.

In the experiments for our convection-dominated Navier-Stokes problem above, we saw that the tested W-methods with inexact matrices encounter stability issues for large time steps. Now we want to briefly examine, whether the minimum time step size, from which on some specific W-methods do produce viable solutions, depends on the fineness of the spatial discretization. I.e., we want to see, whether we have to fulfill a CFL-type condition here.

To do that, we study numerical experiments, very similar to the ones presented above. We still look at problem 6.3.1 and we also use the same spatial and temporal discretization techniques that we described at the beginning of this section.

However, for the subsequent experiments, we do not keep the spatial discretization fixed to always having squares with sides of length 2^{-5} as we did above. Instead, we vary the spatial refinement level k between 4, 5, 6, 7 and 8 — with the corresponding side length of the squares, which we use for the Q_2 -elements, being 2^{-k} . In addition, we do not use a maximum length of $T = 1$ for the time interval, but rather lengthen it to $T = 10$ to improve our ability of detecting stability issues in the time-stepping schemes. We do still use the same exact solution (6.3) and set the right-hand side accordingly.

Now we choose the kinematic viscosity ν *just* small enough, so that the tested time-stepping method fails to produce a viable solution for the spatial refinement level 4 (i.e., the squares of the spatial mesh have sides of length 2^{-4}) and the temporal refinement level 0 (i.e., the time step size is $\tau = 0.2 \cdot 2^{-0}$).

Then we increase the temporal refinement level twice, i.e., we decrease the time step size, to see if we obtain viable solutions. After that, we keep the temporal refinement level *fixed* at $k = 2$ but increase the spatial refinement level to see if that leads to stability problems.

We test the method Scholz45W (Scholz45 as W-method with the same stiffness matrix at each time step), one of the better order 2 W-methods, and the method ROS34PRWW (ROS34PRW as W-method with the same stiffness matrix at each time step), the best order 3 W-method when judging by the l^2H^1 -velocity errors obtained from the previous experiments.

Scholz45W with $\nu = 0.04$				
l	k	$l^\infty L^2$ - velocity error	$l^2 H^1$ - velocity error	$l^2 L^2$ - pressure error
4	0	7.257599e+08	3.547295e+10	2.983200e+09
4	1	1.250379e-02	2.737411e-01	2.709361e-01
4	2	3.030283e-03	6.776393e-02	6.460873e-02
5	2	3.016641e-03	6.776589e-02	6.466069e-02
6	2	3.015339e-03	6.776285e-02	6.468649e-02
7	2	3.015238e-03	6.776236e-02	6.469822e-02

Table 6.32: A Navier-Stokes problem with known solution. Testing independence on spatial refinement for Scholz45W: l denotes the spatial refinement level (with the squares of the spatial mesh having sides of length 2^{-l}) and k denotes the temporal refinement level (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

ROS34PRWW with $\nu = 0.02$				
l	k	$l^\infty L^2$ - velocity error	$l^2 H^1$ - velocity error	$l^2 L^2$ - pressure error
4	0	2.109297e+12	1.779763e+14	1.146720e+13
4	1	1.686025e-03	4.829610e-02	2.511194e-01
4	2	1.528703e-04	6.989056e-03	6.168567e-02
5	2	1.583920e-04	8.416585e-03	6.161797e-02
6	2	1.584806e-04	9.071255e-03	6.159949e-02
7	2	1.581447e-04	9.230241e-03	6.159632e-02

Table 6.33: A Navier-Stokes problem with known solution. Testing independence on spatial refinement for ROS34PRWW: l denotes the spatial refinement level (with the squares of the spatial mesh having sides of length 2^{-l}) and k denotes the temporal refinement level (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)

As we can see, even though the problem is chosen in such a way that the experiments are conducted near a maximally admissible time step size, substantially increasing the fineness of the spatial mesh does not require us to further decrease the time step size in order to circumvent stability issues. Furthermore, the $l^\infty L^2$ -velocity errors, the $l^2 H^1$ -velocity errors and the $l^2 L^2$ -pressure errors are not significantly influenced by the fineness of the spatial mesh either.

6.4 Benchmark Flow Around a Circular Obstacle

In this last section on numerical experiments, we finally look at a more realistic problem setup — one, where the exact solution is not known. Our choice is a two dimensional variant of the famous benchmark problem of a flow around a circular obstacle. And as is commonly done, we conduct experiments in which the Reynolds number is such that one expects the von Kármán vortex street to form behind the obstacle.

Since the exact solution for this type of problem is not known and, of course, we still need to somehow assess the quality of our methods, we want the exact setup of our experiments to

be as close as possible to the setup of some experiment that has been well-documented in the scientific literature. That way we can compare the solutions produced by our methods with many other numerical results and ideally a reference solution that has been computed on a very fine mesh and with very small time steps. We opted for the specific setup described and studied in several papers by John and others, with the papers [30] by John, [32] by John, Matthies and Rang and [33] by John and Rang being our main sources.

The problem setup described in those papers is as follows:

Let $S := \{(x, y) \in \mathbb{R}^2 : \|(x, y) - (0.2, 0.2)\|_2 \leq 0.05\}$ and $\Omega := (0, 2.2) \times (0, 0.41) \setminus S$ as illustrated below in the figure 6.17, which we more or less copied from the paper [33]. Furthermore, let $T := 8$, $\nu := 10^{-3}$ and define $g : (0, T] \rightarrow C(\partial\Omega)^2$ by setting

$$g(t) [(x, y)] := \begin{cases} 0.41^{-2} \sin\left(\frac{\pi t}{8}\right) \begin{pmatrix} 6y(0.41 - y) \\ 0 \end{pmatrix} & \text{if } x \in \{0, 2.2\} \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & \text{else} \end{cases}$$

for all $t \in (0, T]$ and $(x, y) \in \partial\Omega$. Using those definitions and the usual $Q(\Omega) := \{p \in L^2(\Omega) : (p, 1)_{L^2} = 0\}$ notation, we now present the problem formulation:

Problem 6.4.1. We seek a $u = (v, p) : [0, T] \rightarrow H^1(\Omega)^2 \times Q(\Omega)$ with

$$\begin{aligned} \frac{dv}{dt}(t) - \nu \Delta v(t) + (v(t) \cdot \nabla) v(t) + \nabla p(t) &= 0 & \text{for all } t \in (0, 1], \\ \operatorname{div} v(t) &= 0 & \text{for all } t \in (0, 1], \\ v(t)|_{\partial\Omega} &= g(t) & \text{for all } t \in (0, 1], \\ v(0) &= 0. \end{aligned}$$

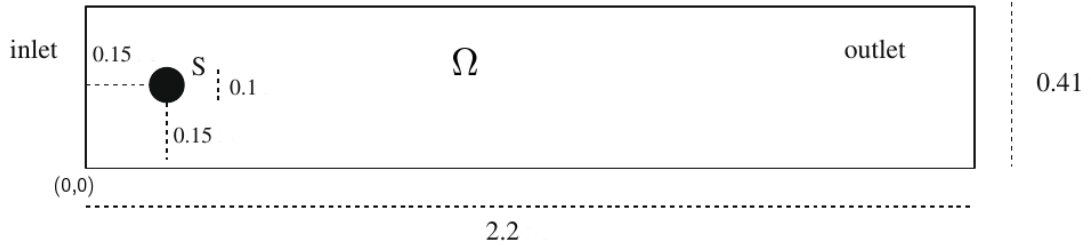


Figure 6.17: Domain Ω for the benchmark flow around a circular obstacle S with diameter 0.1. Image taken directly from [33]

We cover the domain Ω with the mesh illustrated in figure 6.18 and uniformly refine that mesh three more times to arrive at the final mesh that we utilize for our experiments. We employ the same temporal and spatial discretization techniques as those that are described at the beginning of the last section 6.3 - including the LPS-stabilized Q_2/Q_2 -spatial discretization.

This also means, though, that in contrast to our problem setup, which is *exactly* as in the

papers [30, 32, 33], our spatial discretization technique unfortunately differs from the unstabilized Q_2/P_1^{disc} -spatial discretization technique that is used in those papers. In order to obtain comparable results, however, we chose a final mesh that leads to a number of degrees of freedom in our spatial discretization which is similar to the number of degrees of freedom in the spatial discretization described in the papers [30, 32, 33].

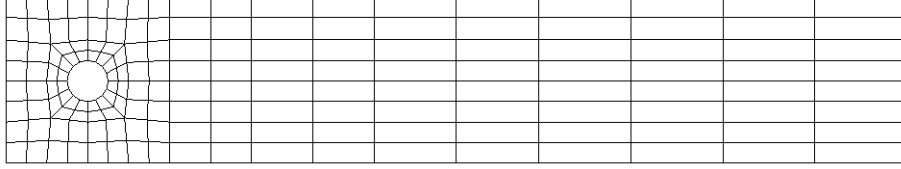


Figure 6.18: Level 1 mesh for the benchmark flow around a circular obstacle

We again test all methods from section 4.6 and the Crank-Nicolson method. For details on the implementation of those methods in fully-discrete algorithms, including the handling of the pressure, the incompressibility condition and the boundary values, we refer to the beginning of the previous section 6.3. In the figures and tables below we denote the Crank-Nicolson method by CN, the ROS2 method with $\gamma = 1 + \frac{1}{\sqrt{2}}$ by ROS2p and the ROS2 method with $\gamma = 1 - \frac{1}{\sqrt{2}}$ by ROS2m.

In our first set of numerical experiments we test the methods from section 4.6 as ROW methods (denoted below with ROW at the end of the name of the method) and as W-methods (denoted below with W at the end of the name of the method) for which we build the stiffness matrix only at the first time step and then use that same matrix as approximate stiffness matrix for *every* subsequent time step. In our second set of experiments, we then test the methods Scholz45 and ROS34PRW also as W-methods for which we build the stiffness matrix at the first step and then update it every 5th/20th/80th step. Those W-method variants will be denoted with W5/W20/W80 at the end of the name of the methods.

Unfortunately, the numerical results produced by the Crank-Nicolson method and all variants of Scholz45 display large unwanted oscillations. A reason why specifically Crank-Nicolson and Scholz45 produce substantial oscillations could be that these methods do not have the best stability properties, as both of them are merely A-stable — while ROS3PW is at least *strongly* A-stable and ROS2 and ROS34PRW are even L-stable. To still obtain usable numerical solutions, we thus employ a post-processing for all numerical solutions obtained from Crank-Nicolson and our Scholz45 variants and indicate those post-processed results with the word *smoothed*.

Our post processing is very simple but produces surprisingly smooth and accurate solutions. If $M + 1 \in \mathbb{N}_{\geq 3}$ is the number of time steps, $h > 0$ describes the mesh size and $\{(v_{m,h}, p_{m,h}) : m \in \{0, \dots, M\}\}$ is a fully-discrete numerical solution with large unwanted oscillations, we calculate a new smoothed numerical solution $\{(\tilde{v}_{m,h}, \tilde{p}_{m,h}) : m \in \{0, \dots, M\}\}$ by setting:

$$\begin{aligned} (\tilde{v}_{h,0}, \tilde{p}_{h,0}) &:= \frac{1}{2} ((v_{h,0}, p_{h,0}) + (v_{h,1}, p_{h,1})), \\ (\tilde{v}_{h,M}, \tilde{p}_{h,M}) &:= \frac{1}{2} ((v_{h,M-1}, p_{h,M-1}) + (v_{h,M}, p_{h,M})), \\ \forall m \in \{1, \dots, M-1\} : (\tilde{v}_{h,m}, \tilde{p}_{h,m}) &:= \frac{1}{4} ((v_{h,m-1}, p_{h,m-1}) + 2(v_{h,m}, p_{h,m}) + (v_{h,m+1}, p_{h,m+1})). \end{aligned}$$

In order to assess the quality of our methods, we calculate, after each time step, a lift coefficient and a drag coefficient of the numerical solution (the smoothed one, of course, for CN and Scholz45) and we also examine a specific pressure difference (see below) at the end time $T = 8$. These are characteristic values for the benchmark flow around a circular obstacle and from the paper [33] we can obtain corresponding reference values that were created by using trusted numerical methods and very fine spatial and temporal discretizations.

At any point in time $t \in [0, T]$ let $(v(t), p(t)) \in H^1(\Omega)^2 \times Q(\Omega)$ denote some velocity/pressure pair, let $n = (n_x, n_y)$ denote the outward pointing *normal* vector of ∂S and let $v_{t_{\partial S}}(t)$ denote the *tangential* component of $v(t)$ at the boundary ∂S of S . The corresponding drag coefficient $c_d(t)$ and lift coefficient $c_l(t)$ are then defined as follows:

$$c_d(t) := 20 \int_{\partial S} \left(\nu n_y \frac{\partial v_{t_{\partial S}}}{\partial n}(t) - n_x p(t) \right) d(\partial S), \quad (6.4)$$

$$c_l(t) := -20 \int_{\partial S} \left(\nu n_x \frac{\partial v_{t_{\partial S}}}{\partial n}(t) + n_y p(t) \right) d(\partial S). \quad (6.5)$$

We want to mention, that in the paper [30] it is argued, that a different method for calculating these coefficients by means of volume integrals rather than line integrals is more accurate and less sensitive to the way the boundary of S is approximated by the mesh. With the software we utilize, however, it is a lot easier to implement the above formulas for the drag and lift coefficients. Hence, we do use the formulas (6.4) and (6.5).

For a given pressure p , the characteristic pressure difference δp is defined as the difference between the pressure just in front of the obstacle S and the pressure just behind the obstacle S , both taken at the end time $t = T = 8$. I.e., we define $\delta p := p(8) [(0.15, 0.2)] - p(8) [(0.25, 0.2)]$.

The time step sizes chosen in the papers [30, 32, 33] range from rather large values such as 0.04 to small values such as 0.00125 to being determined through adaptive time step control. Testing many different time step sizes for this benchmark problem goes beyond the scope of our work, though. Therefore, we chose the time step size $\tau = 0.01$ as a middle ground. However, many of the W-method variants we test do not produce feasible solutions for $\tau = 0.01$. In our first experiments, where we look at ROW methods and only those W-method variants that use the same matrix at each time step, we thus decrease the time step for those W-method variants until we reach one that is small enough to obtain a feasible solution. In our second set of experiments, we keep the time step fixed at $\tau = 0.01$ but increase the update frequency of the stiffness matrix in Scholz45 and ROS34PRW a few times. We then examine which of the tested update frequencies is sufficient to maintain stability and we also show at what time exactly our unstable W-method variants of Scholz45 and ROS34PRW explode.

We first look at the lift coefficients from our experiments with the ROW methods and those W-method variants that use the same matrix at each time step:

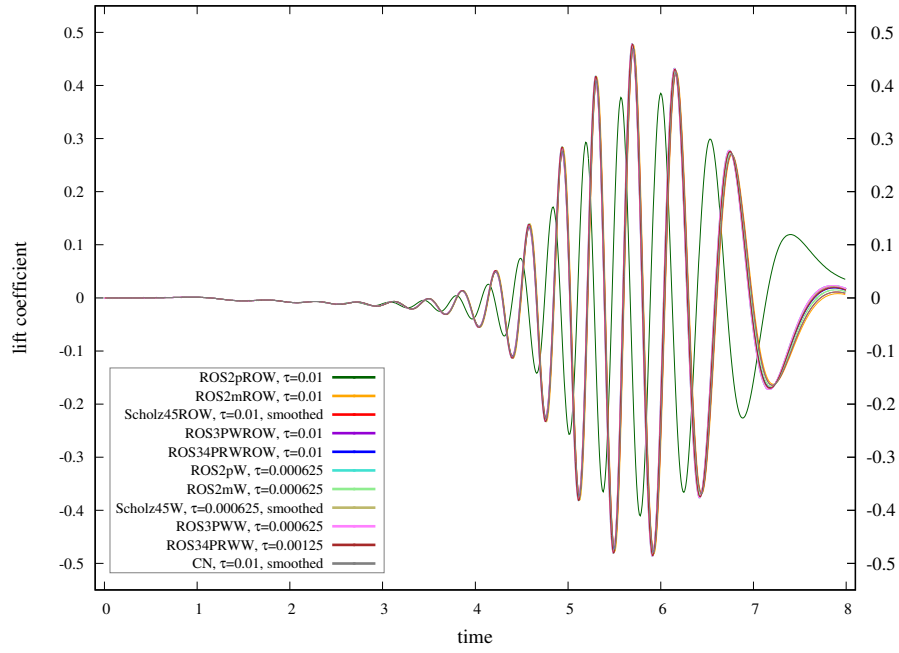


Figure 6.19: Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method

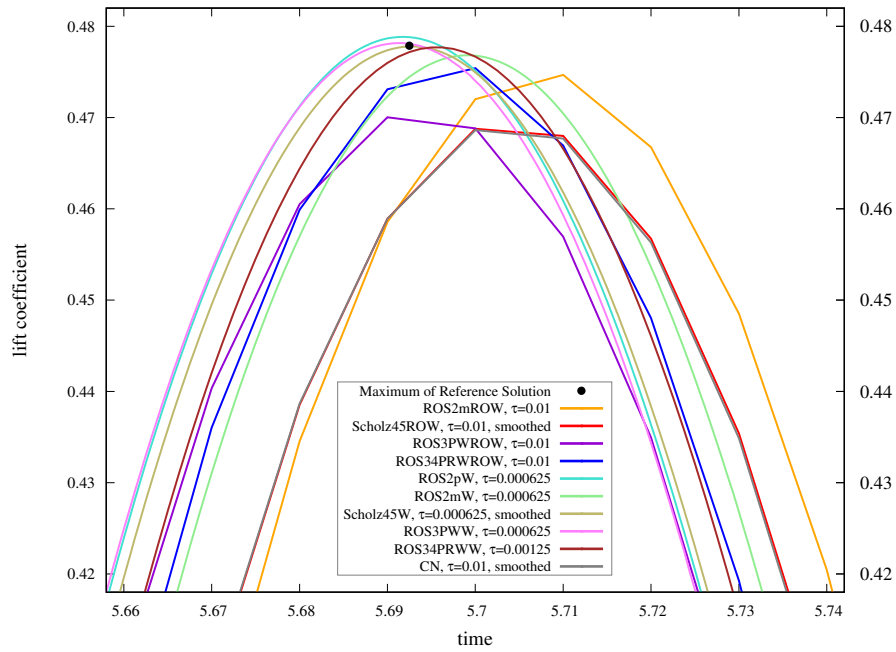


Figure 6.20: Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method — zoomed in near the maximum lift of the reference solution from [33]

As we can see in the zoomed out figure 6.19, the graphs of almost all lift coefficients corresponding to the different tested methods have very similar shapes — with the one exception being the graph of the lift coefficient corresponding to ROS2pROW with $\tau = 0.01$. We do not have the data to also plot a complete reference lift coefficient from the paper [33] or the paper [32]. From the pictures in [32] it becomes clear, though, that the reference solution therein has a lift coefficient whose graph has a shape which is very similar to the shapes seen in our figure 6.19 (excluding the lift coefficient corresponding to ROS2pROW with $\tau = 0.01$, of course).

In the zoomed in figure 6.20 it becomes even more apparent, that most of our tested methods perform very well. The methods with small time steps hit the reference maximum from [33] almost exactly, but the method ROS34PRWROW with $\tau = 0.01$ also comes very close to that reference maximum.

Next, we present the drag coefficients for our first set of benchmark flow experiments:

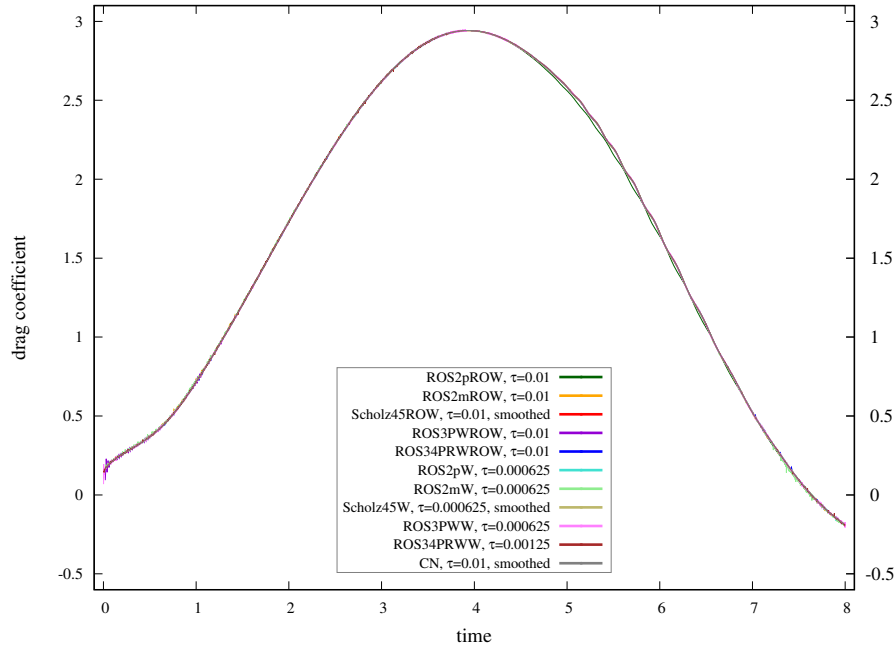


Figure 6.21: Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method

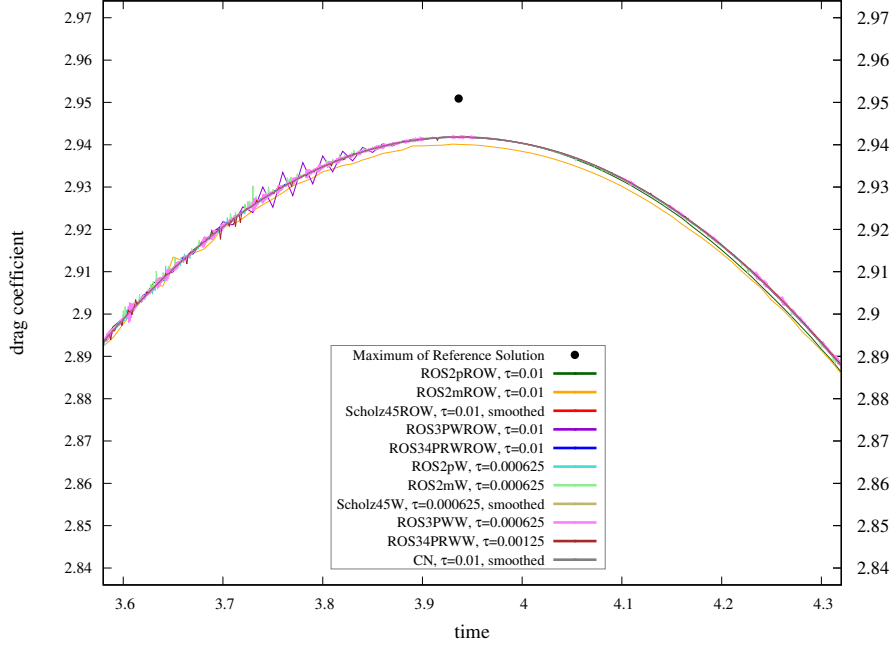


Figure 6.22: Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method — zoomed in near the maximum drag of the reference solution from [33]

First of all, we see that the general shapes of the graphs of the drag coefficients all look very similar and also close to what is shown for a reference solution in [30]. Looking a bit closer, we also notice, though, that the drag coefficients corresponding to many methods display some small oscillations that do not look like they should be present in an exact solution — especially when comparing to the graph of the reference drag from [30]. Remember that we already smoothed the numerical solutions produced by Scholz45 and CN through post processing. The drag coefficients of those numerical solutions only look so smooth because of that post processing — without it they would display extreme oscillations.

The interesting observation is that even in more stable methods these small oscillations are visible. In the zoomed in figure 6.22, we see that ROS3PW (both as ROW method and as W-method with the same matrix at every time step) produces rather large oscillations, as do the tested W-method variants of ROS2p, ROS2m and ROS34PRW. But even the ROW method variants of ROS2p, ROS2m and ROS34PRW show some, albeit very small, oscillations.

We suspect, that the stability — or lack thereof — of the time-stepping methods is not the sole source of these oscillations. This idea is compounded by the observations we make, when we play around with different stabilization parameters for the local projection stabilization (LPS) we use. Increasing a specific stabilization parameter helps to decrease oscillations in the drag coefficient, but also makes it so that the maximum of the drag coefficients decreases even further. Notice that in figure 6.22 the drag coefficient maxima of our tested methods are very similar, even though the methods differ greatly in time step size and in the amount of times a stiffness matrix is computed. On the other hand, the maximum of the reference drag coefficient is significantly larger than all those drag maxima produced by our methods. This makes us think that the spatial discretization

— i.e., Q_2/Q_2 with LPS — and our specific approximation of the boundary of S have a much larger influence on the errors in the drag coefficients than our temporal discretization techniques.

For the sake of completeness, we also list the pressure differences (and the corresponding errors compared to the reference solution from [33]) for the tested ROW methods and W-methods:

	δp	$\delta p - \delta p_{\text{ref}}$	$\left \frac{\delta p - \delta p_{\text{ref}}}{\delta p_{\text{ref}}} \right $
Reference Solution from [33]	-0.11161567	0	0
ROS2pROW, $\tau = 0.01$	-0.106499	5.1167e-03	4.5842e-02
ROS2mROW, $\tau = 0.01$	-0.113482	-1.8663e-03	1.6721e-02
Scholz45ROW, $\tau = 0.01$, smoothed	-0.10936	2.2557e-03	2.0209e-02
ROS3PWROW, $\tau = 0.01$	-0.111793	-1.7733e-04	1.5888e-03
ROS34PRWROW, $\tau = 0.01$	-0.111497	1.1867e-04	1.0632e-03
ROS2pW, $\tau = 0.000625$	-0.111433	1.8267e-04	1.6366e-03
ROS2mW, $\tau = 0.000625$	-0.112012	-3.9633e-04	3.5508e-03
Scholz45W, $\tau = 0.000625$, smoothed	-0.111510	1.0567e-04	9.4673e-04
ROS3PWW, $\tau = 0.000625$	-0.121197	-9.5813e-03	8.5842e-02
ROS34PRWW, $\tau = 0.00125$	-0.111908	-2.9233e-04	2.6191e-03
CN, $\tau = 0.01$, smoothed	-0.110941	6.7467e-04	6.0446e-03

Table 6.34: Benchmark flow around a circular obstacle: pressure differences for the reference solution from [33], several ROW methods and W-methods and the Crank-Nicolson method

Now we present the results for our second set of numerical experiments in which we test many different W-method variants of Scholz45 and ROS34PRW — keeping the time step size fixed at $\tau = 0.01$ but varying the update frequency of the stiffness matrix:

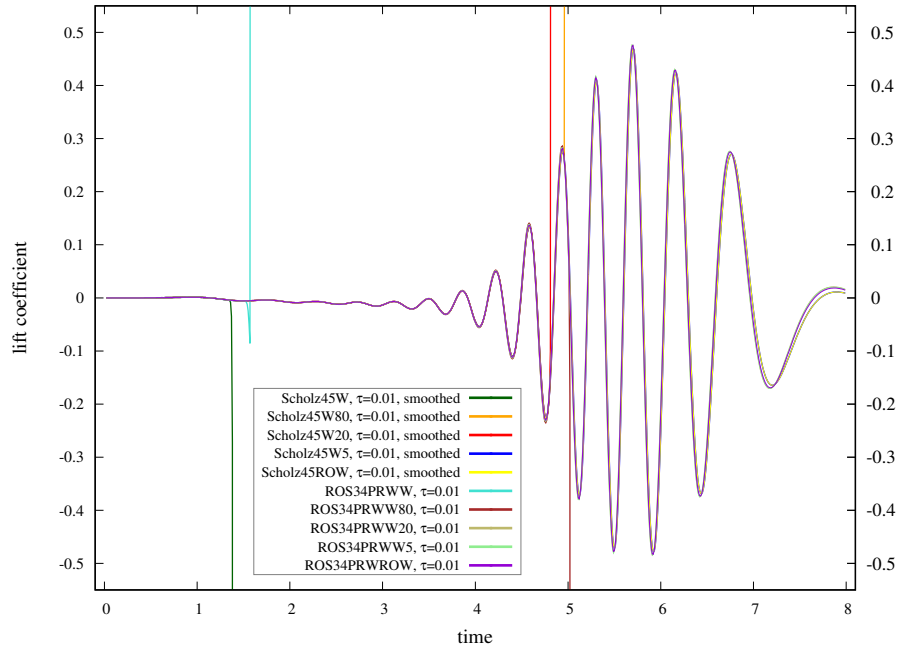


Figure 6.23: Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW

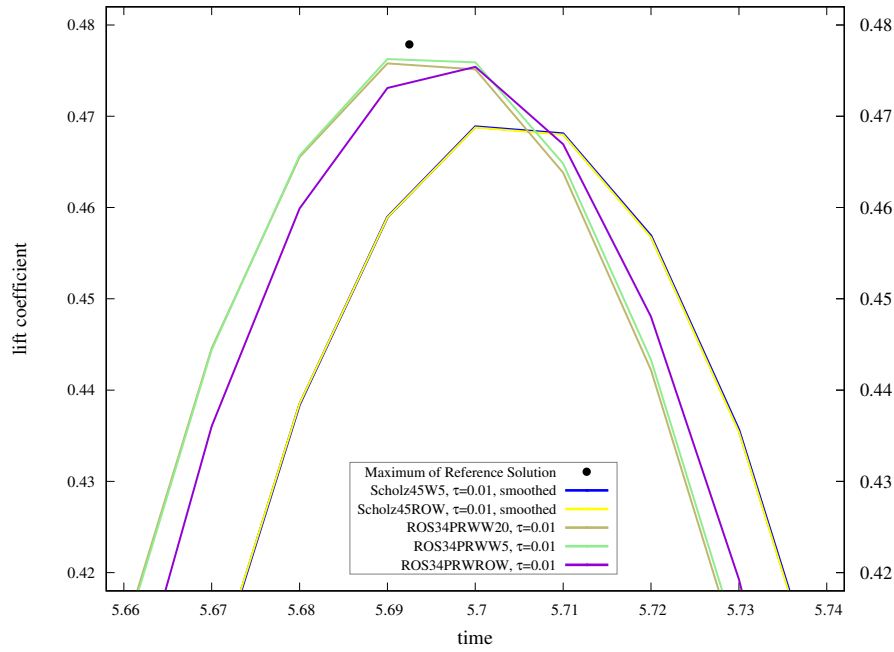


Figure 6.24: Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW — zoomed in near the maximum lift of the reference solution from [33]

The above figure 6.23 nicely shows how updating the stiffness matrix more often increases the point in time at which the tested W-method explodes. Among our tested update frequencies, updating at every step or at every 5th step leads to full solutions for both Scholz45 and ROS34PRW, with ROS34PRW even producing a full solution when the matrix is updated only every 20th step. The other update frequencies lead to exploding numerical solutions.

The graphs of the lift coefficients — excluding the exploding parts, of course — closely resemble what is seen in figure 6.19 further above and in a graph of a reference drag in [32].

In the zoomed in figure 6.24, we see that the lift maxima produced by ROS34PRWW5 and ROS34PRWW20 come *very* close to the lift maximum of the reference solution from [33]. In fact, all tested methods that do not explode provide rather accurate lift coefficients. Regarding these numerical results for the benchmark flow, we thus make the observation that when a W-method does not explode, then the solution it produces is very accurate. Or in other words, it seems as if the greatest issue with W-methods that utilize inexact stiffness matrices is their stability.

The figures 6.25, 6.26 and the table 6.35 below show the corresponding drag coefficients and pressure differences. As in our first set of experiments, we again observe a notable discrepancy between the drag maxima of the tested W-methods and the drag maximum of the reference solution from [33]. And once again, we suspect, that the main source for that discrepancy is likely not the temporal discretization but the stabilized finite elements and our specific approximation of the boundary of S — for some more details on that see the discussion of the drag results for our first set of experiments further above.

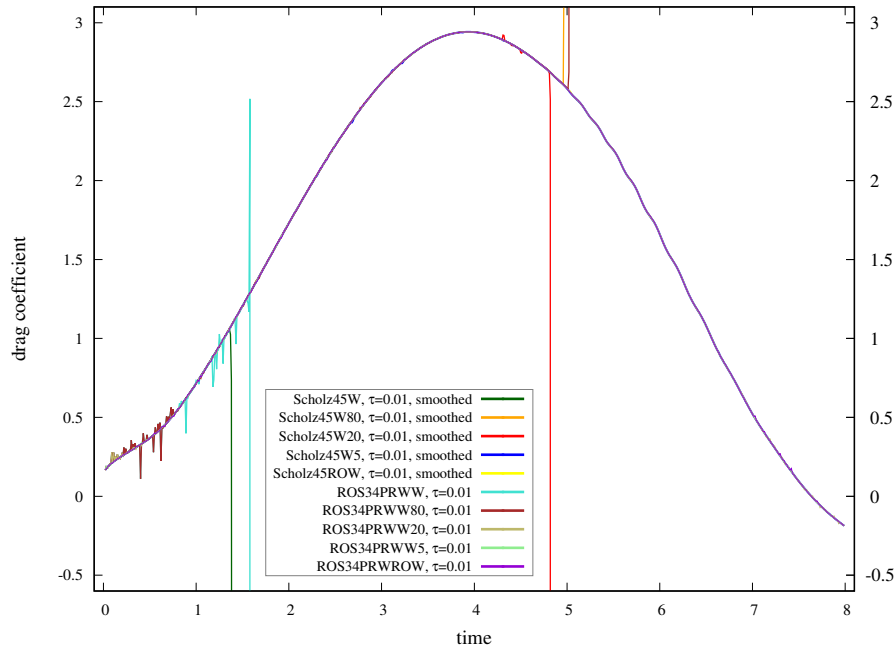


Figure 6.25: Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW

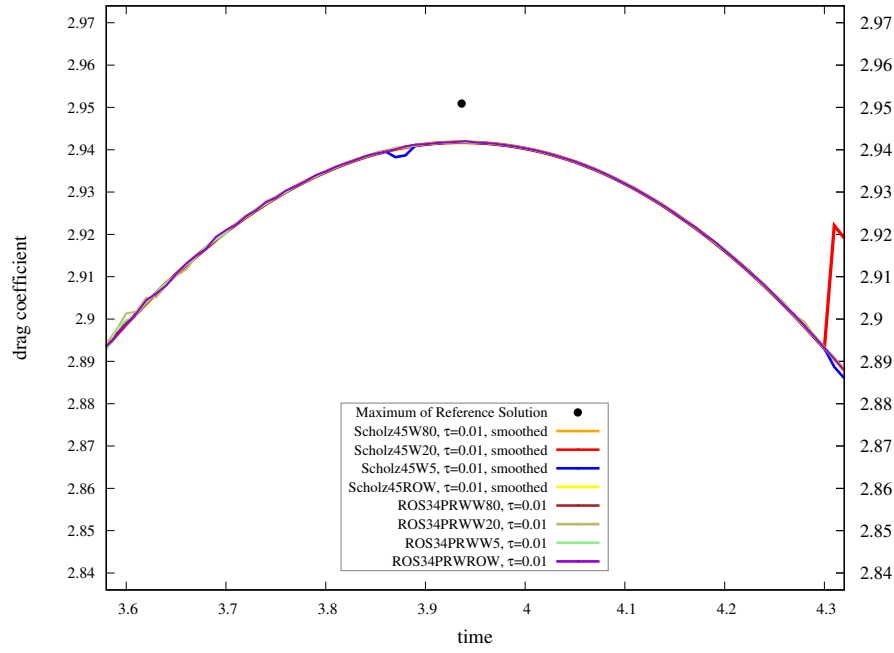


Figure 6.26: Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW — zoomed in near the maximum drag of the reference solution from [33]

	δp	$\delta p - \delta p_{\text{ref}}$	$\left \frac{\delta p - \delta p_{\text{ref}}}{\delta p_{\text{ref}}} \right $
Reference Solution from [33]	-0.11161567	0	0
Scholz45W, $\tau = 0.01$, smoothed	exploded	exploded	exploded
Scholz45W80, $\tau = 0.01$, smoothed	exploded	exploded	exploded
Scholz45W20, $\tau = 0.01$, smoothed	exploded	exploded	exploded
Scholz45W5, $\tau = 0.01$, smoothed	-0.109832	1.7837e-03	1.5980e-02
Scholz45ROW, $\tau = 0.01$, smoothed	-0.10936	2.2557e-03	2.0209e-02
ROS34PRWW, $\tau = 0.01$	exploded	exploded	exploded
ROS34PRWW80, $\tau = 0.01$	exploded	exploded	exploded
ROS34PRWW20, $\tau = 0.01$	-0.112181	-5.6533e-04	5.0650e-03
ROS34PRWW5, $\tau = 0.01$	-0.121654	-1.0038e-02	8.9937e-02
ROS34PRWROW, $\tau = 0.01$	-0.111497	1.1867e-04	1.0632e-03

Table 6.35: Benchmark flow around a circular obstacle: pressure differences for the reference solution from [33] and the ROW method variants and several W-method variants of Scholz45 and ROS34PRW

This concludes the last set of numerical results we present in this work. Overall, the performance of our tested ROW methods and W-methods seems very promising. The ROW methods are very stable and accurate but require a new stiffness matrix to be build at every time step. The tested W-methods with inexact stiffness matrices are also surprisingly accurate. They do run into stability issues, though — depending on the size of the diffusion coefficient/kinematic viscosity and the update frequency of the stiffness matrix.

The suitability of W-methods with inexact stiffness matrices for the temporal discretization of semilinear parabolic equations thus depends on the answers to multiple questions:

- Does the nonlinearity contain convective terms (less suitable) or only zero-order terms (more suitable)?
- How large (more suitable) or small (less suitable) is the diffusion coefficient/the kinematic viscosity?
- Is it numerically expensive (more suitable) or inexpensive (less suitable) to build the stiffness matrix?
- Is the time step size required to be very small (more suitable) already for accuracy reasons or would the accuracy provided by larger time steps (less suitable) be sufficient?

Regarding these W-methods, there are also quite a few things still to be investigated in detail. Namely, how beneficial in terms of saved computing time it actually is to build the stiffness matrix less often, and also how the stiffness matrix should best be approximated or how often it should be updated. One might also try to find indicators, like embedded error estimators, that signal when to update the matrix.

7 Summary, Conclusion and Outlook

The purpose of this work was to investigate the suitability of Rosenbrock-type methods (specifically Rosenbrock-Wanner (ROW) methods and W-methods) for the temporal discretization of parabolic partial differential equations (PDEs) with convective terms and lower-order nonlinearities — such as the incompressible Navier-Stokes equations. The motivation to try out these time-stepping methods stemmed from the hope that computing time might be saved by building the stiffness matrix less often and solving fewer linear systems than in implicit Runge-Kutta methods without losing much stability or accuracy.

While the application of ROW methods (exact Fréchet derivatives are used) to the Navier-Stokes equations and other parabolic equations had already been extensively studied — both in theory and in numerical experiments — the related W-methods (Fréchet derivatives may be approximated) seemed to not have been explored in that much detail yet.

Our approach was two fold: Firstly, we carefully examined if and how existing theory for the semi-discretization in time by Rosenbrock-type methods was applicable to the incompressible Navier-Stokes equations and some convection-diffusion-reaction equations. Secondly, we conducted various numerical experiments with different ROW methods and W-methods and the aforementioned equations to test the performance of the methods and the utility of the semi-discrete error bounds.

Summary and Conclusion

Regarding the theoretical examination, we restricted ourselves to studying in detail only the *semi*-discretization in time — rather than any specific *fully*-discrete algorithms. This, however, let us at least gauge the usefulness of Rosenbrock-type methods for the problems that interest us — independent of any spatial discretization.

Much of the existing literature on the semi-discretization in time by Rosenbrock-type methods uses an abstract framework of Gelfand triples, analytic semigroups and sectorial operators. Examples are the lecture notes [37] by Lang and the papers [40, 42, 41] by Lubich and Ostermann — all of those we extensively used in our work. As such, we were faced with the somewhat intricate task of showing that the parabolic PDEs which interest us could be cast into the before mentioned abstract framework and that they fulfill all requirements of the semi-discrete error bounds in the paper [41].

In chapter 3 we prepared our proof of the applicability of the error bounds in the paper [41]. We first established the needed concepts of sectorial operators, analytic semigroups and semilinear

parabolic equations (SPEs). Then we demonstrated in detail — carefully handling any asymmetries introduced by convective terms — that our newly defined convection-diffusion operator, the Stokes operator and our newly defined Oseen operator can all be viewed as sectorial operators which can be extended to bounded operators on certain Gelfand triples. This extending of sectorial operators to bounded operators — inspired by the beginning of section 2. in [42] — would be very useful when later showing that linearizing the incompressible Navier-Stokes equations and some convection-diffusion-reaction equations leads to operators that have all the properties required for the semi-discrete error bounds.

Overall, embedding our convection-diffusion-reaction equations and the incompressible Navier-Stokes equations into the above described abstract framework — albeit being somewhat technical — did not pose too many difficulties and seemed worthwhile to do. It gives us the opportunity to use elegant concepts such as analytic semigroups and also makes it possible to treat these parabolic PDEs almost like ordinary differential equations (ODEs) with values in Hilbert spaces. Moreover, there seems to exist a lot of literature on these topics, both in pure mathematics (see for example the earlier works [35] by Kato and [27] by Henry or the later book [73] by Yagi), and also in numerics (see for example the above mentioned works by Lang, Lubich and Ostermann).

At the end of chapter 3, we also supplied, for the sake of completeness and without proofs, some known results on the existence and uniqueness of solutions to semilinear parabolic equations. Furthermore, we carefully restated Henry’s (see example 3.8 in [27]) identification of the incompressible Navier-Stokes equations as a semilinear parabolic equation.

In the following chapter 4, we first took a detour to the field of ODEs. This let us more easily describe the idea behind Rosenbrock-type methods, such as ROW methods and W-methods, which is to linearize a nonlinear equation and then treat the linearization implicitly and the nonlinear error term explicitly — with the ROW methods requiring an exact Jacobian as the linearization, while W-methods, which can be viewed as a generalization of ROW methods, allow for approximations of the Jacobian. Another reason why we started with ODEs was that the properties of these methods, also when applied to semilinear (Hilbert space) equations, are usually still described with terminology and concepts that are used for ODEs, such as (classical) convergence order and the various standard notions of stability — this, of course, is justified by the above mentioned close relation between SPEs and ODEs.

We then demonstrated how Rosenbrock-type methods are used for the semi-discretization in time of SPEs and formulated two theorems with semi-discrete error bounds — one for ROW methods and one for W-methods. Those theorems were constructed as slightly simplified versions of results from the paper [41]. Despite our simplifications, there were still numerous requirements on the method and especially on the equation that one had to check.

And that is what we did in the second to last section of chapter 3: We showed that our aforementioned theorems for ROW methods and W-methods are applicable to the incompressible Navier-Stokes equations and some variants of convection-diffusion-reaction equations.

For the W-method theorem, this was a bit more involved because there were additional requirements to fulfill: Namely, we also had to check whether the operators with which we wanted to approximate the exact Fréchet derivatives fulfilled all requirements stated in the convergence theorem. This was especially difficult for the incompressible Navier-Stokes equations, for which we were only able to *conjecture* but not *firmly prove* that approximating the exact Fréchet de-

rivatives with only occasional updates is covered by the theory. Nevertheless, we were able to validate the requirements of our semi-discrete convergence theory for quite a few different choices of the approximative operator in W-methods — both for the incompressible Navier-Stokes equations and some variants of convection-diffusion-reaction equations.

We also touched briefly on the somewhat optimistic assumption that the exact solutions in our problems have relatively high temporal regularity. We more or less justified that assumption by describing the phenomenon of parabolic smoothing. However, we also mentioned some non-smooth semi-discrete error bounds.

In the short chapter 5 we gave a rudimentary introduction to our means of discretizing in space: the Finite Element Method (FEM). We did not, though, describe in detail our fully-discrete algorithm nor did we derive any fully-discrete error bounds, as that was beyond the scope of this work. Rather, we referred to results from the lecture notes [37] by Lang, in which the author constructs fully-discrete estimates for ROW methods in a framework very similar to ours.

Finally, we performed a variety of numerical experiments with a few different methods (the parameters are given in 4.6), each of which we tested both as ROW method and as W-method with inexact Fréchet derivatives. The results were presented in chapter 6, and in most experiments we chose the solution and then set the initial and boundary conditions as well as the right-hand side accordingly. We conducted these experiments with known solution for convection-diffusion problems, reaction-diffusion problems and incompressible Navier-Stokes problems. Each problem was examined both with moderate and with small diffusion coefficient.

The experiments showed that if in W-methods the convective term was omitted from the stiffness matrix, or only occasionally updated, then those W-methods would encounter stability issues, i.e., they would not produce feasible solutions for large time steps — we only observed this for small diffusion coefficient, though. When they did produce feasible solutions, most W-methods with inexact matrices performed very well, often not losing too much accuracy (sometimes even maintaining it) compared to their ROW method counterparts.

The observed numerical convergence orders, both for the ROW methods and the W-methods with inexact matrices, always reached at least the orders given by the semi-discrete error bounds, for some methods and time step sizes they were even larger. Especially noteworthy in that regard is the method ROS34PRW, which — when used as a W-method that omits the convective term from the stiffness matrix — displayed for some problems a *fractional* numerical convergence order between 2 and 3 — with the method's classical ODE order being 3 (both as ROW method and as W-method) and its semi-discrete W-method order being predicted as *at least* 2 for small enough time steps.

This observation suggests that the semi-discrete error bound 4.5.2 for W-methods, which we obtained from the paper [41], might not be sharp for some methods and problems. That is not too much of a surprise, though, as in that paper (and in our somewhat simplified version of the theorem therein) the W-methods are only required to have at least classical order 2 — and ROS34PRW has classical order 3 also when used as a W-method with inexact Jacobians.

Finally, we also conducted experiments with an incompressible Navier-Stokes problem for which the exact solution was not known. The setup we chose for our experiments was a variant (see the paper [30] by John and the paper [33] by John and Rang for details) of the well-known benchmark

problem of a two-dimensional flow around a circular obstacle — with the von Kármán vortex street forming behind the obstacle.

Once again, as in the experiments with known solutions and higher Reynolds numbers, we observed that W-methods which used one and the same stiffness matrix at each time step required relatively small time steps to produce feasible solutions — we could, however, decrease the required time step size by updating the stiffness matrix occasionally. The accuracy of the numerical solutions — measured by comparing drag, lift and pressure difference against a reference solution that was calculated in a similar setting — appeared to be good for ROW methods and overall very promising for all W-method variants that produced feasible solutions.

Outlook

This brings us to the areas of research and unanswered questions related to this work that, in our opinion, deserve further investigation.

Firstly, we think that it should be made clear how much computing time is actually saved by solving fewer linear systems, and, especially, how much is saved by building the stiffness matrix less often. For example, one could conduct numerical experiments with W-methods (varying the update frequency of the stiffness matrix) and Runge-Kutta methods at different refinement levels of the spatial discretization and then, most importantly, plot the accurateness of the numerical solution against *CPU time*.

In addition, it would certainly make sense to improve our setup for the benchmark flow around an obstacle by using formulas for calculating lift and drag coefficients and a spatial discretization that are not just very similar but *exactly equal* to the formulas and discretizations that are used in [33] — the paper from which we obtained the reference values. That way, the performance of Rosenbrock-type methods could certainly be checked much more accurately.

When the cost and benefit of assembling the stiffness matrix is adequately assessed, one could then seek suitable strategies for deciding how often — or even at which time steps exactly — a new matrix should be built in W-methods. In our experiments we merely tested W-method variants for which the matrix was updated every 5th/20th/80th time step — with those numbers being picked rather arbitrarily. One could, for example, try to design indicators, such as error estimators which are based on embedded lower order methods, that signal when to build a new matrix. In this context it should also be interesting to compare W-methods that use approximate stiffness matrices with Runge-Kutta methods that use inexact Newton methods to solve the nonlinear systems.

Regarding the analytical side of things, we were, unfortunately, unable to simplify the rather technical requirements and proofs of the existing semi-discrete error bounds — achieving that would definitely be desirable. As already explained above, some results from our numerical experiments furthermore suggest that the semi-discrete order 2 error bound 4.5.2 for W-methods might not be sharp for some problems and methods with higher classical order. Hence, finding and proving sharp semi-discrete error estimates similar to the ones in the paper [41] but for higher order W-methods could be another worthwhile objective for the future.

Moreover, seeking and proving precise *fully*-discrete error estimates for algorithms that involve ROW methods or W-methods and variants of the finite element method is certainly a major challenge but — judging by what we learned in our research for this work — might be worth the effort. Lastly, one might also consider widening the theoretical and experimental research of this work to other problems, such as magnetohydrodynamics.

List of Figures

2.1	A fixed control volume V with boundary ∂V and outward pointing unit normal field n . f denotes the flux.	8
3.1	A sector $S_{a,\phi}$ with opening angle $\gamma := 2\pi - 2\phi$	20
3.2	A contour Γ around the spectrum $\sigma(\mathbb{A})$ of a sectorial operator \mathbb{A} with sector $S_{a,\phi}$	22
3.3	Two nested Sectors $S_{a,\phi} \subset S_{a,\phi'}$ and a parabolic area in $\mathbb{C} \setminus S_{a,\phi'}$ which contains the spectrum $\sigma(\mathbb{D})$ of a convection-diffusion operator \mathbb{D}	28
6.1	A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	122
6.2	A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	125
6.3	A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	127
6.4	A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	129
6.5	A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	134
6.6	A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	137
6.7	A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	139
6.8	A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	141
6.9	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^\infty L^2$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	147
6.10	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 H^1$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	150

6.11	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 L^2$ -pressure error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	152
6.12	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	154
6.13	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	157
6.14	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 L^2$ -pressure error versus temporal refinement level k (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	159
6.15	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution — testing occasional updating of the stiffness matrix in Scholz45 and ROS34PRW: $l^2 H^1$ -velocity error versus temporal refinement level k (with $5 \cdot 2^k$ time steps and the time step size $0.2 \cdot 2^{-k}$ in the experiments)	161
6.16	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution — testing occasional updating of the system matrix in Scholz45 and ROS34PRW: $l^2 L^2$ -pressure error versus temporal refinement level k (with $5 \cdot 2^k$ time steps and the time step size $0.2 \cdot 2^{-k}$ in the experiments)	162
6.17	Domain Ω for the benchmark flow around a circular obstacle S with diameter 0.1. Image taken directly from [33]	165
6.18	Level 1 mesh for the benchmark flow around a circular obstacle	166
6.19	Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method	168
6.20	Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method — zoomed in near the maximum lift of the reference solution from [33]	168
6.21	Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method	169
6.22	Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for ROW methods, W-methods (same stiffness matrix at every time step) and the Crank-Nicolson method — zoomed in near the maximum drag of the reference solution from [33]	170
6.23	Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW	172
6.24	Benchmark flow around a circular obstacle: lift coefficient c_l versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW — zoomed in near the maximum lift of the reference solution from [33]	172
6.25	Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW	173
6.26	Benchmark flow around a circular obstacle: drag coefficient c_d versus time t for the ROW method variants and several W-method variants of Scholz45 and ROS34PRW — zoomed in near the maximum drag of the reference solution from [33]	174

List of Tables

6.1	A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	121
6.2	A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	121
6.3	A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	124
6.4	A convection-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	124
6.5	A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	126
6.6	A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	126
6.7	A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	128
6.8	A convection-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	129
6.9	A convection-diffusion problem with known solution. Testing independence on spatial refinement for Scholz45W and ROS34PRWW: l denotes the spatial refinement level (with the squares of the spatial mesh having sides of length 2^{-l}) and k denotes the temporal refinement level (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	131
6.10	A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	133
6.11	A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	134
6.12	A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	136
6.13	A reaction-diffusion problem with diffusion coefficient $\nu = 1$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	136
6.14	A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	138

6.15	A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	138
6.16	A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	140
6.17	A reaction-diffusion problem with diffusion coefficient $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	140
6.18	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^\infty L^2$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	146
6.19	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^\infty L^2$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	146
6.20	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 H^1$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	149
6.21	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 H^1$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	150
6.22	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 L^2$ -pressure errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	151
6.23	A Navier-Stokes problem with kinematic viscosity $\nu = 1$ and known solution: $l^2 L^2$ -pressure errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	151
6.24	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	153
6.25	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^\infty L^2$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	154
6.26	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	156
6.27	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	156
6.28	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 L^2$ -pressure errors for ROW methods and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	158
6.29	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 L^2$ -pressure errors for W-methods and the Crank-Nicolson method and all time step sizes $\tau = 0.2 \cdot 2^{-k}$ with $k \in \{0, \dots, 8\}$	158
6.30	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: $l^2 H^1$ -velocity errors for Scholz45 and ROS34PRW (system matrix update only every 80th/20th/5th time step), with temporal refinement level $k \in \{0, \dots, 8\}$, $5 \cdot 2^k$ time steps and time step size $0.2 \cdot 2^{-k}$	160

6.31	A Navier-Stokes problem with kinematic viscosity $\nu = 10^{-4}$ and known solution: l^2L^2 -pressure errors for Scholz45 and ROS34PRW (system matrix update only every 80th/20th/5th time step), with temporal refinement level $k \in \{0, \dots, 8\}$, $5 \cdot 2^k$ time steps and time step size $0.2 \cdot 2^{-k}$	162
6.32	A Navier-Stokes problem with known solution. Testing independence on spatial refinement for Scholz45W: l denotes the spatial refinement level (with the squares of the spatial mesh having sides of length 2^{-l}) and k denotes the temporal refinement level (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	164
6.33	A Navier-Stokes problem with known solution. Testing independence on spatial refinement for ROS34PRWW: l denotes the spatial refinement level (with the squares of the spatial mesh having sides of length 2^{-l}) and k denotes the temporal refinement level (with the time step size being $\tau = 0.2 \cdot 2^{-k}$)	164
6.34	Benchmark flow around a circular obstacle: pressure differences for the reference solution from [33], several ROW methods and W-methods and the Crank-Nicolson method	171
6.35	Benchmark flow around a circular obstacle: pressure differences for the reference solution from [33] and the ROW method variants and several W-method variants of Scholz45 and ROS34PRW	174

Bibliography

- [1] R. ADAMS AND J. FOURNIER, *Sobolev Spaces*, vol. 140 of Pure and Applied Mathematics, Academic Press, Boston, MA, 2003.
- [2] H. D. ALBER, *Elliptische partielle Differentialgleichungen*. https://www2.mathematik.tu-darmstadt.de/fbereiche/analysis/pde/teaching/Skripten_Alber/elli.pdf. accessed 22-April-2019.
- [3] H. W. ALT, *Linear Functional Analysis. An Application-Oriented Introduction*, Universitext, Springer, London, 2016. <https://doi.org/>.
- [4] H. AMANN, *Nonhomogeneous Linear and Quasilinear Elliptic and Parabolic Boundary Value Problems*, in : H.-J. Schmeisser, H. Triebel (eds) Function Spaces, Differential Operators and Nonlinear Analysis, vol. 133 of Teubner-Texte zur Mathematik, Vieweg+Teubner, Wiesbaden, 1993, pp. 9–126. https://doi.org/10.1007/978-3-663-11336-2_1.
- [5] W. AUZINGER, R. FRANK, AND G. KIRLINGER, *A Note on Convergence Concepts for Stiff Problems*, Computing, 44 (1990), pp. 197–208. <https://doi.org/10.1007/BF02262216>.
- [6] R. BECKER AND M. BRAACK, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, CALCOLO, 38 (2001), pp. 173–199. <https://doi.org/10.1007/s10092-001-8180-4>.
- [7] M. BRAACK, *Numerik gewöhnlicher Differentialgleichungen*, 2017. <http://www.math.uni-kiel.de/angewandte-mathematik/de/lehre-teaching/SkriptNumDglBraackKap1-8-1.pdf>. accessed 24-April-2019.
- [8] M. BRAACK, *Finite Elements I & II*, 2018. <http://www.math.uni-kiel.de/angewandte-mathematik/de/lehre-teaching/FEM-Braack.chapter1-14.pdf>. accessed 17-April-2019.
- [9] M. BRAACK AND E. BURMAN, *Local Projection Stabilization for the Oseen Problem and its Interpretation as a Variational Multiscale Method*, SIAM J. Numer. Anal., 43 (2006), pp. 2544–2566. <https://doi.org/10.1137/050631227>.
- [10] M. BRAACK, E. BURMAN, V. JOHN, AND G. LUBE, *Stabilized finite element methods for the generalized oseen problem*, Comput. Method. Appl. M., 196 (2007), pp. 853–866. <https://doi.org/10.1016/j.cma.2006.07.011>.
- [11] D. BRAESS, *Finite Elemente. Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*, Springer, Berlin, Heidelberg, 2013. <https://doi.org/10.1007/978-3-642-34797-9>.
- [12] R. COURANT, K. FRIEDRICHS, AND H. LEWY, *Über die partiellen Differenzengleichungen der mathematischen Physik*, Math. Ann., 100 (1928), pp. 32–74. <https://doi.org/10.1007/BF01448839>.

-
- [13] G. DAHLQUIST, *A special stability problem for linear multistep methods*, BIT, 3 (1963), pp. 27–43. <https://doi.org/10.1007/BF01963532>.
- [14] M. DAUGE, *Stationary Stokes and Navier–Stokes Systems on Two- or Three-Dimensional Domains with Corners. Part I. Linearized Equations*, SIAM J. Math. Anal., 20 (1989), pp. 74–97. <https://doi.org/10.1137/0520006>.
- [15] P. DEUFLHARD, *Uniqueness Theorems for Stiff ODE Initial Value Problems*, in : D. F. Griffiths and G. A. Watson (eds) Numerical Analysis 1989, Proceedings of the 13th Dundee Conference, Pitman Research Notes in Mathematics Series, Longman Scientific and Technical, 1990, pp. 74–87. <https://nbn-resolving.org/urn:nbn:de:0297-zib-53>. accessed 9-May-2019.
- [16] L. C. EVANS, *Partial Differential Equations. second edition*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, 2010. <http://dx.doi.org/10.1090/gsm/019>.
- [17] R. FRANK, J. SCHNEID, AND C. W. UEBERHUBER, *The Concept of B-Convergence*, SIAM J. Numer. Anal., 18 (1981), pp. 753–780. <https://doi.org/10.1137/0718051>.
- [18] H. FUJITA AND T. KATO, *On the Navier-Stokes Initial Value Problem. I*, Arch. Rational Mech. Anal., 16 (1964), pp. 269–315. <https://doi.org/10.1007/BF00276188>.
- [19] D. FUJIWARA, *Concrete Characterization of the Domains of Fractional Powers of Some Elliptic Differential Operators of the Second Order*, Proc. Japan Acad., 43 (1967), pp. 82–86. <https://doi.org/10.3792/pja/1195521686>.
- [20] GASCOIGNE: THE FINITE ELEMENT TOOLKIT. <http://www.uni-kiel.de/gascoigne/>. accessed 8-May-2019.
- [21] P. GRISVARD, *Caractérisation de Quelques Espaces d’Interpolation*, Arch. Ration. Mech. An., 25 (1967), pp. 40–63. <https://doi.org/10.1007/BF00281421>.
- [22] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, vol. 69 of Classics in Applied Mathematics, SIAM, 2011. <https://doi.org/10.1137/1.9781611972030>.
- [23] J.-L. GUERMOND AND A. SALGADO, *A note on the Stokes operator and its powers*, J. Appl. Math. Comput., 36 (2011), pp. 241–250. <https://doi.org/10.1007/s12190-010-0400-0>.
- [24] M. HAASE, *The Functional Calculus for Sectorial Operators and Similarity Methods*, Dissertation, Universität Ulm, 2003. https://www.uni-ulm.de/fileadmin/website_uni_ulm/mawi.inst.020/share/diss/Haase2003.disfinal.pdf. accessed 22-April-2019.
- [25] E. HAIRER AND G. WANNER, *On the Butcher group and general multi-value methods*, Computing, 13 (1974), pp. 1–15. <https://doi.org/10.1007/BF02268387>.
- [26] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*, vol. 14 of Springer Series in Computational Mathematics, Springer, Berlin, Heidelberg, 1996. <https://doi.org/10.1007/978-3-642-05221-7>.
- [27] D. HENRY, *Geometric Theory of Semilinear Parabolic Equations*, vol. 840 of Lecture Notes in Mathematics, Springer, Berlin, Heidelberg, 1981. <https://doi.org/10.1007/BFb0089647>.

- [28] M. HERMANN, *Numerik gewöhnlicher Differentialgleichungen. Anfangs- und Randwertprobleme*, De Gruyter, Berlin, Boston, 2009. <https://www.degruyter.com/view/product/230513>. accessed 24-April-2019.
- [29] W. H. HUNSDORFER, *Stability and B-Convergence of Linearly Implicit Runge-Kutta Methods*, Numer. Math., 50 (1986), pp. 83–95. <https://doi.org/10.1007/BF01389669>.
- [30] V. JOHN, *Reference values for drag and lift of a two-dimensional time-dependent flow around a cylinder*, Int. J. Numer. Meth. Fluids, 44 (2004), pp. 777–788. <https://doi.org/10.1002/flid.679>.
- [31] V. JOHN, *Finite Element Methods for Incompressible Flow Problems*, Springer, 2016. <https://doi.org/10.1007/978-3-319-45750-5>.
- [32] V. JOHN, G. MATTHIES, AND J. RANG, *A comparison of time-discretization/linearization approaches for the incompressible Navier–Stokes equations*, Comput. Method. Appl. M., 195 (2006), pp. 5995–6010. <https://doi.org/10.1016/j.cma.2005.10.007>.
- [33] V. JOHN AND J. RANG, *Adaptive time step control for the incompressible navier–stokes equations*, Comput. Method. Appl. M., 199 (2010), pp. 514–524. <https://doi.org/10.1016/j.cma.2009.10.005>.
- [34] P. KAPS, *Modifizierte Rosenbrockmethoden der Ordnung 4, 5 und 6 zur numerischen Integration steifer Differentialgleichungen*, Dissertation, (1977).
- [35] T. KATO, *Perturbation Theory for Linear Operators*, vol. 132 of Classics in Mathematics, Springer, Berlin, Heidelberg, 1995. <https://doi.org/10.1007/978-3-642-66282-9>.
- [36] D. KUZMIN, *A Guide to Numerical Methods for Transport Equations*, 2010. <http://www.mathematik.uni-dortmund.de/~kuzmin/Transport.pdf>. accessed 17-April-2019.
- [37] J. LANG, *Adaptive Multilevel Solution of Nonlinear Parabolic PDE Systems. Theory, Algorithm, and Applications*, vol. 16 of Lecture Notes in Computational Science and Engineering, Springer, Berlin, Heidelberg, 2001. <https://doi.org/10.1007/978-3-662-04484-1>.
- [38] J. LANG AND D. TELEAGA, *Towards a Fully Space-Time Adaptive FEM for Magnetoquasistatics*, IEEE Trans. Magn., 44 (2008), pp. 1238–1241. <https://doi.org/10.1109/TMAG.2007.914837>.
- [39] S. LARSSON, *Nonsmooth Data Error Estimates With Applications To The Study Of The Long-Time Behavior Of Finite Element Solutions Of Semilinear Parabolic Problems*, Preprint, 1992:36 (1992). <http://www.math.chalmers.se/~stig/papers/preprints.html>. accessed 2-May-2019.
- [40] C. LUBICH AND A. OSTERMANN, *Runge-Kutta Methods for Parabolic Equations and Convolution Quadrature*, Math. Comput., 60 (1993), pp. 105–131. <https://doi.org/10.1090/S0025-5718-1993-1153166-7>.
- [41] C. LUBICH AND A. OSTERMANN, *Linearly implicit time discretization of non-linear parabolic equations*, IMA J. Numer. Anal., 15 (1995), pp. 555–583. <https://doi.org/10.1093/imanum/15.4.555>.
- [42] C. LUBICH AND A. OSTERMANN, *Runge-Kutta Approximation of Quasi-Linear Parabolic Equations*, Math. Comput., 64 (1995), pp. 601–627. <https://doi.org/10.1090/S0025-5718-1995-1284670-0>.

-
- [43] A. OSTERMANN, *Stability of W-methods with applications to operator splitting and to geometric theory*, Appl. Numer. Math., 42 (2002), pp. 353–366. [https://doi.org/10.1016/S0168-9274\(01\)00160-X](https://doi.org/10.1016/S0168-9274(01)00160-X).
 - [44] A. OSTERMANN AND M. ROCHE, *Runge-Kutta Methods for Partial Differential Equations and Fractional Orders of Convergence*, Math. Comput., 59 (1992), pp. 403–420. <https://doi.org/10.1090/S0025-5718-1992-1142285-6>.
 - [45] A. OSTERMANN AND M. ROCHE, *Rosenbrock methods for partial differential equations and fractional orders of convergence*, SIAM J. Numer. Anal., 30 (1993), pp. 1084–1098. <https://doi.org/10.1137/0730056>.
 - [46] A. OSTERMANN AND M. THALHAMMER, *Non-smooth data error estimates for linearly implicit Runge-Kutta methods*, IMA J. Numer. Anal., 20 (2000), pp. 167–184. <https://doi.org/10.1093/imanum/20.2.167>.
 - [47] S. PANKAVICH AND N. MICHALOWSKI, *A Short Proof of Increased Parabolic Regularity*, Electron. J. Diff. Equ., 205 (2015), pp. 1–9. <https://arxiv.org/abs/1502.01773>.
 - [48] A. PROTHERO AND A. ROBINSON, *On the Stability and Accuracy of One-Step Methods for Solving Stiff Systems of Ordinary Differential Equations*, Math. Comput., 28 (1974), pp. 145–162. <https://doi.org/10.1090/S0025-5718-1974-0331793-2>.
 - [49] J. RANG, *A New Stiffly Accurate Rosenbrock-Wanner Method for Solving the Incompressible Navier-Stokes Equations*, in : R. Ansorge, H. Bijl, A. Meister, T. Sonar (eds) Recent Developments in the Numerics of Nonlinear Hyperbolic Conservation Laws, Notes on Numerical Fluid Mechanics and Multidisciplinary Design, Springer, Berlin, Heidelberg, 2013, pp. 301–315. https://doi.org/10.1007/978-3-642-33221-0_18.
 - [50] J. RANG, *Improved traditional Rosenbrock–Wanner methods for stiff ODEs and DAEs*, J. Comput. Appl. Math., 286 (2015), pp. 128–144. <https://doi.org/10.1016/j.cam.2015.03.010>.
 - [51] J. RANG, *The Prothero and Robinson example: Convergence studies for Runge–Kutta and Rosenbrock–Wanner methods*, Appl. Numer. Math., 108 (2016), pp. 37–56. <https://doi.org/10.1016/j.apnum.2016.04.012>.
 - [52] J. RANG AND L. ANGERMANN, *New Rosenbrock W-Methods of Order 3 for Partial Differential Algebraic Equations of Index 1*, BIT Numer. Math., 45 (2005), pp. 761–787. <https://doi.org/10.1007/s10543-005-0035-y>.
 - [53] J. RANG AND L. ANGERMANN, *New Rosenbrock methods of order 3 for PDAEs of index 2*, in Proceedings of Equadiff 11, Comenius University Press, Bratislava, 2007, pp. 385–394. <http://eudml.org/doc/219949>.
 - [54] T. RICHTER, *Numerische Methoden der Strömungsmechanik*, 2015. <https://www.math.uni-magdeburg.de/~richter/pdf-files/numerik3.pdf>. accessed 22-April-2019.
 - [55] T. RICHTER, A. SPRINGER, AND B. VEXLER, *Efficient numerical realization of discontinuous Galerkin methods for temporal discretization of parabolic problems*, Numer. Math., 124 (2013), pp. 151–182. <https://doi.org/10.1007/s00211-012-0511-7>.
 - [56] H. H. ROSENBRICK, *Some general implicit processes for the numerical solution of differential equations*, Comput. J., 5 (1963), pp. 329–330. <https://doi.org/10.1093/comjnl/5.4.329>.

- [57] E. ROTHE, *Zweidimensionale parabolische Randwertaufgaben als Grenzfall eindimensionaler Randwertaufgaben*, Math. Ann., 102 (1930), pp. 650–670. <https://doi.org/10.1007/BF01782368>.
- [58] S. SCHOLZ, *Order Barriers for the B-Convergence of ROW Methods*, Computing, 41 (1989), pp. 219–235. <https://doi.org/10.1007/BF02259094>.
- [59] B. SCHWEIZER, *Partielle Differentialgleichungen. Eine anwendungsorientierte Einführung*, Springer, Berlin, Heidelberg, 2013. <https://doi.org/10.1007/978-3-642-40638-6>.
- [60] B. SCHWEIZER, *Evolutionsgleichungen. Lineare und Semilineare Halbgruppentheorie*, 2018. <http://www.mathematik.uni-dortmund.de/lisi/schweizer/Skripte/evolution2018.pdf>. accessed 17-April-2019.
- [61] F. SCHWITZER, *W-methods for semilinear parabolic equations*, Appl. Numer. Math., 18 (1995), pp. 351–366. [https://doi.org/10.1016/0168-9274\(95\)00062-Y](https://doi.org/10.1016/0168-9274(95)00062-Y).
- [62] T. STEIHAUG AND A. WOLFBRANDT, *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations*, Math. Comput., 33 (1979), pp. 521–534. <https://doi.org/10.1090/S0025-5718-1979-0521273-8>.
- [63] C. STOVER, *Sobolev Embedding Theorem*. <http://mathworld.wolfram.com/SobolevEmbeddingTheorem.html>. accessed 8-May-2019.
- [64] K. STREHMEL AND R. WEINER, *Linear-implizite Runge-Kutta-Methoden und ihre Anwendung*, vol. 127 of Teubner-Texte zur Mathematik, Vieweg+Teubner, Wiesbaden, 1992. <https://doi.org/10.1007/978-3-663-10673-9>.
- [65] K. STREHMEL, R. WEINER, AND M. BÜTTNER, *Order results for Rosenbrock type methods on classes of stiff equations*, Numer. Math., 59 (1991), pp. 723–737. <https://doi.org/10.1007/BF01385806>.
- [66] K. STREHMEL, R. WEINER, AND H. PODHAISKY, *Numerik gewöhnlicher Differentialgleichungen. Nichtsteife, steife und differential-algebraische Gleichungen*, Vieweg+Teubner, Wiesbaden, 2012. <https://doi.org/10.1007/978-3-8348-2263-5>.
- [67] I. TELEAGA AND J. LANG, *Higher-Order Linearly Implicit One-Step Methods for Three Dimensional Incompressible Navier-Stokes Equations*, Studia Babes-Bolyai Matematica, 53 (2008), pp. 109–121. <http://www.cs.ubbcluj.ro/~studia-m/2008-1/teleaga.pdf>. accessed 8-May-2019.
- [68] R. TEMAM, *Navier-Stokes Equations. Theory and Numerical Analysis*, North-Holland, Amsterdam, New-York, Oxford, 1977.
- [69] R. TEMAM, *Navier-Stokes Equations and Nonlinear Functional Analysis*, vol. 66 of CBMS-NSF Regional Conference Series in Applied Mathematics, Siam, 1995. <https://doi.org/10.1137/1.9781611970050>.
- [70] J. VERWER, E. J. SPEE, J. G. BLOM, AND W. HUNSDORFER, *A Second-Order Rosenbrock Method Applied to Photochemical Dispersion Problems*, SIAM J. Sci. Comput., 20 (1999), pp. 1456–1480. <https://doi.org/10.1137/S1064827597326651>.

- [71] WIKIPEDIA CONTRIBUTORS, *Derivation of the navier–stokes equations* — *Wikipedia, the free encyclopedia*. https://en.wikipedia.org/w/index.php?title=Derivation_of_the_Navier%E2%80%93Stokes_equations&oldid=858161413, 2018. [Online; accessed 17-April-2019].
- [72] A. WOLFBRANDT, *A study of Rosenbrock processes with respect to order conditions and stiff stability*, Doctoral Thesis, (1977). <http://hdl.handle.net/2077/15007>.
- [73] A. YAGI, *Abstract Parabolic Evolution Equations and their Applications*, Springer Monographs in Mathematics, Springer, Berlin, Heidelberg, 2010. <https://doi.org/10.1007/978-3-642-04631-5>.

Acknowledgements

I want to thank my supervisor Prof. Dr. Malte Braack for giving me the opportunity of pursuing a doctor's degree in his research group, for always being open to questions and pleasant to work with, and for granting me large freedom in choosing the direction of my research.

Furthermore, I thank all developers of Gascoigne 3D — the finite element toolkit that I used for my numerical experiments — for creating a fast and powerful piece of software and for trying their best in making that software expandable and accessible for new users.

In this context I especially thank Utku Kaya for patiently helping me tame the beast that is Gascoigne 3D and for often lifting my spirits with his optimistic and happy attitude.

I also want to deeply thank Michael Hauptmann for his long-lasting companionship during this endeavor and for the many stimulating discussions about a multitude of interesting subjects.

For proofreading this work, I thank Simon Taylor and again Michael Hauptmann.

Last but not least, I thank my parents, my sister and all my friends for the emotional support during these demanding but fulfilling last years.

Erklärung

Hiermit versichere ich,

- I. dass diese diese Abhandlung - abgesehen von der Beratung durch meinen Betreuer Malte Braack - nach Inhalt und Form meine eigene ist,
- II. dass diese Arbeit an keiner anderen Stelle bereits veröffentlicht, bzw. zur Veröffentlichung eingereicht wurde,
- III. dass diese Arbeit unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft entstanden ist
- IV. und dass mir kein akademischer Grad entzogen wurde.

Ort, Datum

Unterschrift (Leon Schramm)